

Escuela Politécnica Superior

19
20

Trabajo fin de grado

Difusión de la información en grafos: dependencias de la estructura



Sara Sanz Lucio

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C/ Francisco Tomás y Valiente nº 11

**UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR**



Grado en Ingeniería informática

TRABAJO FIN DE GRADO

**Difusión de la información en grafos:
dependencias de la estructura**

Autor: Sara Sanz Lucio

Tutor: Simone Santini

julio 2020

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de la propiedad intelectual.

La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. del Código Penal*).

DERECHOS RESERVADOS

© 9 de Julio de 2020 por UNIVERSIDAD AUTÓNOMA DE MADRID

Francisco Tomás y Valiente, n.º 1

Madrid, 28049

Spain

Sara Sanz Lucio

Difusión de la información en grafos: dependencias de la estructura

Sara Sanz Lucio

C\ Francisco Tomás y Valiente N.º 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

*A mis padres y a Antonio,
por su apoyo incondicional*

*Social media is the ultimate equaliser.
It gives a voice and a platform to anyone willing to engage.*

Amy Jo Martin

PREFACIO

En la última década las redes sociales han alcanzado el protagonismo en nuestra vida diaria. Hoy en día juegan un rol muy relevante no solo en el ámbito del entretenimiento sino también en la generación de opinión. Su alcance actual las convierte en un mecanismo de difusión tanto de información como de desinformación.

En concreto, eventos recientes como las elecciones presidenciales de Estados Unidos de 2016 han propiciado el interés en el estudio de la difusión de la información en las redes debido al impacto que tuvieron en su desarrollo. Con motivo de la aparición de estudios en los últimos años que señalan la importancia de ciertas características estructurales de las redes en la velocidad de difusión, el objetivo de este trabajo es determinar cuáles de estas propiedades afectan en mayor medida a su velocidad de transmisión. Se ha prestado especial atención a aquellas características que provocan la ralentización de este proceso, con el consiguiente impacto en la contención de la información.

Sara Sanz Lucio

AGRADECIMIENTOS

En primer lugar me gustaría agradecer a la Escuela Politécnica Superior y todos sus profesores por la formación recibida en estos últimos cuatro años. En particular agradecer a mi tutor Simone Santini por su guía, sus comentarios y su apoyo durante el desarrollo de este trabajo.

Quisiera también agradecer a mis padres, Alberto y Mamen, por su apoyo en todo momento. Y por supuesto no quiero olvidarme de mi hermano Álvaro, así como del resto de mi familia. Su apoyo ha sido muy importante para que escriba estas líneas. Quiero hacer también una mención especial a Antonio que ha estado a mi lado en todo momento.

RESUMEN

Este trabajo de Fin de Grado ha tenido como objetivo analizar la velocidad de difusión de la información y de la formación de opiniones en las redes sociales y el papel que juega su estructura en este proceso. Para la representación de las redes se han utilizado grafos aleatorios empleando tanto el modelo de Barabási como el de Leskovec para su generación. En lo referente a la difusión de información se ha aplicado el modelo de contacto aleatorio consistente en que en cada iteración cada nodo contacta uno de sus vecinos elegido al azar y, si este vecino contiene la información, el nodo también la consigue.

Para replicar este proceso de difusión se han empleado las variables β y δ , que representan la probabilidad de que una noticia se difunda y se olvide respectivamente. En cuanto a la formación de opiniones, se han considerado los modelos SI (*Susceptible-Infected*) y SIS (*Susceptible-Infected-Susceptible*) derivados del estudio de la propagación de infecciones, así como el modelo de difusión de desinformación que emplea el *fact checking* para la verificación de la veracidad de la información.

Con el fin de entender qué características ralentizan el tiempo de propagación, de forma equivalente a contener la difusión de información en la red, se ha examinado la repercusión que supone realizar ciertas modificaciones sobre las características estructurales de estos grafos.

Dado que la identificación del conjunto de usuarios con tal impacto ha resultado un problema NP completo en estudios similares, se ha empleado una heurística aproximada que elige aquellos nodos con mayor *clustering* y grado como los más influyentes. En concreto se ha observado la variación que implica la eliminación de los nodos con mayor coeficiente de *clustering* y mayor grado sobre la velocidad de la difusión. Esta velocidad ha resultado dependiente en gran medida de estas características estructurales, en especial de aquellos nodos con mayor grado, ya que estos cuentan con una probabilidad de infección potencial mayor al tener más vecinos, por lo que si se les elimina del proceso de propagación el tiempo necesario para alcanzar N infectados aumenta de forma significativa. Esta variación se ha manifestado de forma más pronunciada en el modelo de Leskovec ya que este genera grafos cuyos nodos tienen grados sustancialmente más altos que el modelo de Barabási.

Este estudio se puede extrapolar a la contención de noticias falsas en las redes sociales, ya que al identificar a los usuarios con más influencia en la difusión de estas noticias, se puede controlar su propagación por la red aplicando diferentes medidas de *fact checking* que reconozcan esta información como falsa y regulen su retransmisión, en especial por parte de aquellos usuarios que por su cantidad de conexiones presentan más influencia como se ha mostrado en este estudio. Sin embargo, la implantación de medidas de este tipo se debe estudiar cuidadosamente debido a la posibilidad de interferir con la libertad de expresión de los usuarios en las redes.

PALABRAS CLAVE

Redes sociales, grafos, propiedades estructurales, centralidad, clustering, difusión de información, velocidad de difusión, noticias falsas

ABSTRACT

This work analyzes the rate of information diffusion and the formation of opinions in social networks and the role that network structure plays in this process. For the representation of the networks, random graphs have been used using both the Barabási and Leskovec models for their generation. Information diffusion has been modeled using the random contact model: each node contacting one of its randomly chosen neighbors and, if this neighbor contains the information, the node also gets it with a certain probability.

The local propagation depends mainly on two parameters, β and δ , which represent the probability that a news item will be broadcast and forgotten respectively. Opinion spread has been modeled using models from mathematical epidemiology: the SI (Susceptible-Infected) and SIS (Susceptible-Infected-Susceptible) models derived from the study of the spread of infections, as well as the disinformation dissemination model that uses fact checking to verify the veracity of the information.

In order to understand what characteristics slow down the propagation time, equivalent to containing the dissemination of information on the network, the impact of making certain modifications on the structural characteristics of these graphs has been examined.

The identification of the set of k users that has the greatest impact on the propagation velocity is NP-complete, so a rough heuristic has been used instead that selects those nodes with the highest clustering and node degree as the most influential. Specifically, the variation that implies the elimination of nodes with a higher coefficient of clustering and a greater degree of diffusion speed has been observed. Propagation speed has been found to be strongly dependent on these structural characteristics, especially on those nodes with the highest degree, since they have a greater potential infection probability as they have more neighbors, so if they are eliminated from the propagation process, the time required to reach infected N increases significantly. This variation has been more pronounced in the Leskovec model since it generates graphs whose nodes have substantially higher degrees than the Barabási model.

This study can be extrapolated to the containment of false news on social networks, since by identifying the users with the most influence in the dissemination of this news, its spread through the network can be controlled by applying different measures of fact checking that recognize this information as false and regulate its retransmission, especially by those users who, due to their high number of connections, have more influence as shown in this study. However, the implementation of this type of measures should be carefully studied due to the possibility of interfering with the freedom of expression of users on social networks.

KEYWORDS

Social networks, graphs, structural properties, centrality, clustering, information diffusion, diffusion rate, fake news

ÍNDICE

1	Introducción	1
1.1	Motivación	1
1.2	Objetivos	4
1.3	Organización de la memoria	4
2	Tecnologías a utilizar	5
2.1	Grafos	5
2.1.1	Vecindad	6
2.1.2	Distancia y diámetro	6
2.1.3	Coeficiente de <i>clustering</i>	6
2.2	Las redes sociales como grafos	7
2.2.1	Distribución <i>power law</i>	8
2.2.2	Propiedad de <i>clustering</i>	8
2.3	Modelos de grafos aleatorios	9
2.3.1	Modelo de Barabási	10
2.3.2	Modelo de Leskovec	11
2.4	Modelos de difusión	13
2.4.1	Modelo epidemiológico	13
2.4.2	Modelo de difusión de desinformación en las redes sociales	15
3	Diseño y desarrollo	17
3.1	Velocidad de difusión	17
3.2	Formación de opiniones	18
3.2.1	Modelo SI	18
3.2.2	Modelo SIS	19
3.2.3	Modelo de difusión de desinformación	19
3.3	Medidas	20
3.3.1	Medidas según β	20
3.3.2	Medidas según β y δ	20
3.4	Modificaciones de grafos	20
4	Pruebas y resultados	23
4.1	Características de los grafos generados	23
4.2	Análisis cualitativo de la difusión	26

4.3 Velocidad de difusión siguiendo el modelo epidemiológico SI	29
4.4 Velocidad de difusión siguiendo el modelo epidemiológico SIS	31
4.5 Control de la difusión	32
4.5.1 Efecto de <i>fact-checking</i>	32
4.5.2 Eliminación de los nodos con más vecinos	33
4.5.3 Eliminación de los nodos con más <i>clustering</i>	35
4.5.4 Eliminación de los nodos con menos vecinos y menos <i>clustering</i>	36
5 Conclusiones	37
Bibliografía	40
Definiciones	41
Acrónimos	43

LISTAS

Lista de algoritmos

Lista de códigos

Lista de cuadros

Lista de ecuaciones

2.1	Representación matemática de un grafo simple no dirigido	5
2.2	Representación matemática de un grafo simple dirigido	5
2.3	Coeficiente de <i>clustering</i> en grafos dirigidos	7
2.4	Coeficiente de <i>clustering</i> en grafos no dirigidos	7
2.5	Ecuación diferencial del modelo SI	14
2.6	Ecuación diferencial del modelo SIS	14
2.7	Estado de equilibrio del modelo SIS	14
2.8	Función de difusión de los nodos creyentes en el modelo de difusión de desinformación	15
2.9	Función de defusión de los nodos <i>fact-checkers</i> en el modelo de difusión de desinformación	15
3.1	Propiedad de Markov	17
3.2	Probabilidad de difusión	18
3.3	Variable τ a partir de β y δ	19
3.4	Probabilidad de pasar a estado de creyente y de <i>fact checker</i> respectivamente	19
4.1	Características cualitativas del proceso de difusión	26
4.2	Características cualitativas del proceso de difusión	26
4.3	Características cualitativas del proceso de difusión	26
4.4	Características cualitativas del proceso de difusión	27
4.5	Características cualitativas del proceso de difusión	27
4.6	Características cualitativas del proceso de difusión	27
4.7	Umbral de <i>fact checking</i> a partir del cual desaparece la noticia falsa del grafo	33

Lista de figuras

2.1	Estados de los modelos SI y SIS	14
2.2	Estados del modelo de difusión de desinformación	15
4.1	Representación de la distribución de grado	24
4.2	Representación de la distribución de grado tomando ejes logarítmicos	24
4.3	Representación de la distribución de <i>clustering</i> en los grafos	25
4.4	Velocidad de difusión en el modelo de Barabási empleando el modelo de difusión SI hasta alcanzar el 50 % de los nodos	28
4.5	Velocidad de difusión en el modelo de Barabási empleando el modelo de difusión SI hasta alcanzar el 40 % de los nodos	28
4.6	Velocidad de difusión en el modelo de Barabási empleando el modelo de difusión SI hasta alcanzar el 30 % de los nodos	29
4.7	Velocidad de difusión en el modelo de Leskovec empleando el modelo de difusión SI hasta alcanzar el 40 % de los nodos	29
4.8	Velocidad de difusión en los modelos de Barabási y Leskovec empleando el modelo de difusión SI hasta alcanzar el 50 % de los nodos	30
4.9	Velocidad de difusión en el modelo de Barabási empleando el modelo de difusión SIS hasta alcanzar el 50 % de los nodos	31
4.10	Velocidad de difusión en el modelo de Leskovec empleando el modelo de difusión SIS hasta alcanzar el 50 % de los nodos	31
4.11	Velocidad de difusión en el modelo de Barabási empleando el modelo de difusión de desinformación hasta alcanzar el 50 % de los nodos	32
4.12	Velocidad de difusión en el modelo de Leskovec empleando el modelo de difusión de desinformación hasta alcanzar el 50 % de los nodos	32
4.13	Velocidad de difusión eliminando los nodos con más vecinos en el modelo de Barabási empleando el modelo de difusión SIS hasta alcanzar el 50 % de los nodos	34
4.14	Velocidad de difusión eliminando los nodos con más vecinos en el modelo de Leskovec empleando el modelo de difusión SIS hasta alcanzar el 50 % de los nodos	34
4.15	Velocidad de difusión eliminando los nodos con mayor coeficiente de <i>clustering</i> en el modelo de Barabási empleando el modelo de difusión SIS hasta alcanzar el 50 % de los nodos	35
4.16	Velocidad de difusión eliminando los nodos con mayor coeficiente de <i>clustering</i> en el modelo de Leskovec empleando el modelo de difusión SIS hasta alcanzar el 50 % de los nodos	35
4.17	Velocidad de difusión eliminando los nodos con menos vecinos en el modelo de Barabási empleando el modelo de difusión SIS hasta alcanzar el 50 % de los nodos	36

Lista de tablas

Lista de cuadros

INTRODUCCIÓN

1.1. Motivación

El estudio de las redes sociales resulta de gran relevancia dada su influencia en la transmisión de información hoy en día. Las redes sociales juegan un papel muy importante en nuestra vida, han cambiado nuestra forma de comunicarnos aprovechando el alcance de internet para crear comunidades online, uniendo a usuarios de todo el mundo con intereses comunes. Si a día de hoy (Abril 2020) el alcance de internet es casi universal, con 4500 millones de usuarios, las redes sociales, con 3800 millones, no se quedan muy lejos. Prácticamente el 60 % de la población mundial ya está en línea y las últimas tendencias sugieren que para mediados de 2020 más de la mitad también utilizará las redes sociales según Simon Kemp, analista jefe de DataReportal [1]. La facilidad para acceder a la red y compartir contenido rápidamente es lo que las hace tan interactivas y lo que propicia su rápida propagación, un fenómeno ahora conocido como viralidad.

Un factor clave es su instantaneidad en contraposición a los tiempos de publicación programados de los medios tradicionales, así como la posibilidad de interacción en ambos sentidos que solo es posible en las redes. Es precisamente esta capacidad de poder relacionarse lo que hace que tengan una influencia que ningún otro medio tiene: el poder interactuar con la fuente brinda una sensación de cercanía y fiabilidad que aporta credibilidad al mensaje. Como consecuencia cualquier información que provenga de estas fuentes será mucho más aceptada y difundida por el resto de usuarios. Es este fenómeno el que ha permitido la aparición de la figura del influencer.

Aunque todos los usuarios tengan cierta capacidad de influencia, en las redes sociales los usuarios con más conexiones son considerados más influyentes y una pieza clave en la difusión de la información ya que alcanzan un gran número de usuarios de forma directa. Se podría decir que los influencers son las celebridades de esta generación, tienen la capacidad de crear tendencia y de convencer a un posible cliente de consumir y recomendar un producto. Sin embargo, la combinación del alcance e instantaneidad de las redes y de figuras con tanta influencia también supone riesgos. Se trata de un entorno en el que noticias falsas y estafas (los llamados “bulos”) proliferan con facilidad, explotando la ignorancia de las personas. Este hecho es fácilmente aprovechable por usuarios u organizaciones con

intenciones maliciosas en lo que ahora conocemos como fake news.

Este fenómeno consiste en la creación y divulgación de noticias elaboradas con el fin de inducir a los lectores a creer información manipulada. Es este propósito lo que las hace especialmente peligrosas: en vista de la dificultad que supone detectarlas, estas se producen con distintos objetivos tales como el beneficio económico y político. La propagación de este tipo de noticias ha cobrado más importancia en los últimos años, ya que ha tenido un impacto visible en grandes eventos como las elecciones estadounidenses de 2016. Además, la naturaleza generalmente escandalosa y llamativa de estas noticias motiva que estas se difundan con mayor rapidez a través de las redes. De hecho, en un estudio de 2018 [2], al comparar la difusión de noticias verdaderas y falsas en Twitter se comprobó que aquellas con información falsa se propagan de forma mucho más rápida y extensa. Estas noticias no solo se difunden a mayor velocidad sino que se caracterizan por seguir un proceso de difusión viral. En concreto se observó que la información relativa a política alcanza un conjunto mayor de usuarios siendo más viral que el resto de categorías. Al calcular el modelo de probabilidad se encontró que es un 70 % más probable que un usuario retuitee información falsa frente a información verdadera. Además cabe notar que esto no se debe completamente a las diferencias estructurales de los elementos de la red, ya que se observó que los usuarios que difunden estas noticias tenían significativamente menos seguidores, menos antigüedad en la red y en general una menor probabilidad de ser usuarios verificados y activos en esta. Por ello, se estudió también el comportamiento de los usuarios que motivaba esta divulgación, en concreto la repercusión de la novedad de las noticias en este proceso. La información novedosa no solo capta la atención de los seres humanos sino que además incentiva su propagación debido a su carácter valioso tanto desde una perspectiva teórica relativa a la información como social en cuanto al status que otorga el tener información que no es ampliamente conocida aún. De esta forma se concluyó que es principalmente el comportamiento de los usuarios el que motiva que se compartan de forma más extendida los rumores falsos, por lo que este se debe tener en cuenta también a la hora de desarrollar políticas de contención de esta información.

Las elecciones presidenciales estadounidenses de 2016 son uno de los acontecimientos que mejor ejemplifican el alcance y el peso de las noticias falsas sobre eventos de gran importancia e impacto. Los meses previos a las elecciones las redes se plagaron de noticias falsas relacionadas con los candidatos, despuntando entre ellas un rumor que aseguraba que la candidata demócrata a la presidencia Hillary Clinton y su jefe de gabinete John Podesta dirigían una red de pedofilia desde el sótano de una pizzería de Washington DC. Este rumor estaba supuestamente respaldado por los correos electrónicos de campaña que habían sido interceptados recientemente a causa de un ataque de phishing. Aunque pueda parecer una teoría conspirativa un tanto excéntrica, esta cobró tal relevancia que no tardó en convertirse en tendencia en las redes bajo el hashtag ‘pizzagate’ en Twitter. Este se volvió viral cuando miles de usuarios comenzaron a utilizarlo para tuitear lo que consideraban pruebas tanto apoyando como desmintiendo la conspiración. Sin embargo lo que realmente llamó la atención de este suceso es que una gran cantidad de estos tuits tenían su origen fuera de Estados Unidos, en concreto detec-

tando grandes focos en Chipre, Vietnam y República Checa. Además se observó que algunas de las cuentas que los retuiteaban con más frecuencia eran en realidad bots [3]. Por desgracia, esta situación tuvo trágicas consecuencias cuando un seguidor de esta teoría apareció en la pizzería mencionada con un rifle de asalto exigiendo que se le enseñase el sótano, sin llegar a causar heridos. Las investigaciones realizadas tras el desarrollo de las elecciones sugirieron que esta influencia extranjera jugó un gran papel en el transcurso de las elecciones, con el Centro Nacional de Inteligencia apuntando directamente a la Federación de Rusia y al régimen de Vladimir Putin como posibles culpables.

Como hemos expuesto, en la actualidad las redes sociales no solo son un factor crítico en la difusión de información, sino que además tienen la capacidad de movilizar a la gente. Otros eventos de gran importancia como los levantamientos de la Primavera Árabe también han sido fuertemente influenciados por la inmediatez con la que se comparte la información ya que se considera que la implicación de estas plataformas amplificó la velocidad con la que se desarrollaron los eventos [4].

Así como las redes sociales son un elemento clave para la difusión de noticias, hemos visto que su propagación no es un indicativo de su calidad y veracidad. La pandemia que el mundo está viviendo actualmente lo ejemplifica ya que es el entorno ideal para la proliferación de la desinformación. A lo largo de estos meses hemos visto declaraciones como que “respirar el aire caliente del secador podría tratar la COVID-19”¹ hacerse virales a través de las redes. Esta situación social, política y económica extraordinaria ha creado un clima de incertidumbre en la población, haciendo a esta notablemente propensa a creer y difundir noticias falsas. La propia esencia de las redes promueve esta conducta, ya que su diseño prioriza la cantidad sobre la calidad del contenido, en este caso en forma de engagement por parte de los usuarios. De hecho, según un estudio [5], la motivación de los usuarios a la hora de retuitear también forma parte del problema. Se ha identificado que mientras discutir, mostrar aprobación, entretener o llamar la atención son algunas de las motivaciones de un usuario al compartir, la veracidad de sus argumentos no se encuentra entre ellas.

Estas plataformas han demostrado, una vez más, su alcance, con millones de personas recurriendo a ellas para obtener las últimas noticias y mantenerse en contacto con seres queridos durante el confinamiento. Efectivamente, Twitter ha experimentado un incremento del 23 % en su uso diario en los tres primeros meses de este año con respecto a los últimos tres de 2019 [6]. La propia respuesta de la población ante este brote se está viendo alterada por las redes sociales: con la humillación pública como consecuencia de comportamientos considerados como inapropiados por quienes se consideran “justicieros virtuales”. Este tipo de conductas son el resultado del poder social que otorgan las redes, que no necesariamente tiene por qué desembocar en actitudes positivas. Pero estas críticas no solo afectan al ciudadano medio, sino que el propio gobierno está expuesto a un mayor escrutinio al estar sujeto a la opinión de miles de personas a un clic de distancia. Es esto lo que ha llevado al gobierno español a tomar medidas con el fin de regular toda información relativa a la COVID-19 y su gestión, con la intención de evitar noticias falsas y manipulación. Esta declaración ha alarmado a muchos al ser

¹ Cabe notar que este trabajo se ha realizado durante la pandemia producida por la COVID-19

un recordatorio de la censura en tiempos en los que las redes sociales son el exponente máximo de libertad de expresión.

Se trata de una medida que ha generado un gran debate social, y es de esperar que este debate siga en el futuro, dado que se trata de medidas que demuestran como ciertos conceptos de base para nuestras sociedades, como el de “prensa”, están cambiando, y como en el futuro habrá probablemente que renegociar nuestras exigencias de libertad de expresión con la difusión de una información cuanto más fiable posible.

1.2. Objetivos

Como se ha expuesto, el estudio de la propagación de noticias en las redes sociales es un tema actual de mucha importancia tanto desde el punto de vista sociológico como de la arquitectura de las redes. No obstante, en este trabajo nos vamos a centrar en concreto en la influencia que las características estructurales tienen sobre el proceso de difusión. A saber, se busca entender qué características ralentizan el tiempo de propagación, de forma equivalente a contener la difusión de información en la red. La búsqueda del conjunto de usuarios con tal impacto en el proceso de propagación se supone un problema Nondeterministic Polynomial Time (NP) Completo, ya que existen estudios sobre problemas similares que han llegado a esta conclusión [7].

Por lo tanto, en los próximos capítulos se estudiará la velocidad de propagación de las noticias en las redes, y cómo esta varía al realizar ciertos cambios en las características de las mismas. Entender el efecto que tienen estas propiedades en el tiempo de propagación es vital para poder contrastar la difusión de noticias falsas, ya que como se expone en el estudio antes mencionado [2], estas se difunden a una velocidad mucho mayor que aquellas con información verdadera, lo que en un futuro próximo es pertinente controlar.

1.3. Organización de la memoria

El resto de este trabajo está organizado de la siguiente manera. En el capítulo 2 se explicarán los grafos y cómo se utilizan estos para modelar redes sociales, así como sus características. También se expondrán los modelos que se van a usar para generar estos grafos y simular la difusión de información en estos. A continuación, en el capítulo 3 se explicará el modelo de velocidad de difusión desarrollado, exponiendo cómo se forman las opiniones en los nodos según los distintos modelos de difusión y justificando las modificaciones a realizar en la estructura de los grafos. Finalmente, en el capítulo 4 se expondrán los resultados obtenidos con este modelo, mostrando las características de los grafos generados e interpretando las mediciones obtenidas de las distintas modificaciones realizadas en estos, y en el último capítulo se expresarán las conclusiones obtenidas.

TECNOLOGÍAS A UTILIZAR

2.1. Grafos

Un grafo es una estructura matemática empleada para modelar muchos problemas de manera gráfica por medio de dos elementos básicos: los vértices o nodos y las conexiones entre ellos llamadas aristas.

Un grafo G es un par $G = (V, E)$ siendo V el conjunto de sus vértices y E el de sus aristas. Estas constan de pares de vértices x, y adyacentes que pueden ser ordenados o no dependiendo de si el grafo es dirigido, es decir, si la relación entre los vértices tiene un sentido definido (x, y) , o no, en el caso de relaciones simétricas $\{x, y\}$. Asumiremos siempre que el grafo no contiene self-loop, es decir, por cada $x \in V$, $(x, x) \notin E$.

De esta forma, un grafo simple no dirigido es un par $G = (V, E)$, con

$$\begin{aligned} V &= \{1, \dots, n\}, \\ E &\subseteq \{\{x, y\} \mid (x, y) \in V^2 \wedge x \neq y\} \end{aligned} \tag{2.1}$$

Mientras que un grafo simple dirigido se caracteriza porque sus aristas son pares ordenados de vértices tal que $G = (V, E)$ siendo

$$\begin{aligned} V &= \{1, \dots, n\}, \\ E &\subseteq \{(x, y) \mid (x, y) \in V^2 \wedge x \neq y\} \end{aligned} \tag{2.2}$$

En teoría de grafos existen distintas mediciones para representar sus características estructurales, a continuación se procederá a explicar las más relevantes para este trabajo.

2.1.1. Vecindad

Dado que dos vértices se consideran adyacentes si comparten una arista, se define la vecindad Γ de un vértice u , como las aristas y los vértices adyacentes a este tal que $\Gamma(u) = \{u \in V \mid (v, u) \in E\}$.

Otra forma de caracterizar la vecindad de un vértice u es mediante su grado $\deg(u)$, que expresa el número de aristas que inciden sobre este, expresado tal que $\deg(u) = |\Gamma(u)|$ indica el grado del vértice u . Con el fin de obtener información sobre la estructura de un grafo se emplea esta medida para calcular la distribución de grados en un grafo, de forma que se cuenta la cantidad de nodos que tienen cada grado para obtener la distribución $P_{\deg}(k)$, es decir, la fracción de nodos del grafo con grado k .

2.1.2. Distancia y diámetro

Uno de los problemas más comunes planteados sobre los grafos es el de encontrar el camino más corto entre dos vértices, es decir, el camino conformado por el menor número de vértices y aristas que son necesario atravesar para ir de un nodo u a otro nodo v del grafo. Su resultado permite obtener la distancia existente entre ambos vértices, siendo esta el número de vértices contenidos en este camino.

El diámetro es otra de las mediciones que permiten caracterizar la estructura de un grafo. Este es conocido como el más largo de entre los caminos más cortos existentes, de esta forma, para obtenerlo basta con calcular los caminos más cortos entre cada par de vértices del grafo y elegir el de mayor longitud. Por lo tanto, el diámetro de un grafo se cuantifica como el número de vértices que hay que atravesar para recorrer el grafo desde un vértice hasta otro, excluyendo bucles y backtracking.

2.1.3. Coeficiente de *clustering*

En la mayoría de las redes que se pueden observar en el mundo real y en especial en las redes sociales existe una tendencia a que aparezcan grupos de nodos muy unidos con una gran densidad de conexiones entre ellos. De hecho, su probabilidad tiende a ser mayor que la probabilidad media de que exista una arista establecida aleatoriamente entre dos vértices. Esta propiedad se denomina *cluster* y se mide calculando el grado en que los nodos de un grafo tienden a agruparse, obteniendo el coeficiente de *clustering*. Se pueden considerar dos formas de medir este coeficiente: de manera local y de manera global en el grafo [8].

El coeficiente de *clustering* global se basa en examinar los nodos por tripletes, es decir, tres nodos conectados entre sí por dos o tres aristas. De esta manera, se considera un triángulo como tres tripletes conectados por tres aristas, en otras palabras, tres tripletes “cerrados” en lo referente a sus conexiones. Cada uno de estos tripletes estará centrado en uno de los nodos del triángulo. Luego el coeficiente de *clustering* a nivel global del grafo se calcula como el número de tripletes “cerrados” sobre el número de tripletes total.

De forma análoga, el coeficiente de *clustering* local representa el grado de conectividad existente entre los vecinos de un cierto nodo. Esta medida cuantifica lo cerca que están sus vecinos de ser un *clique* o subgrafo completamente interconectado. Para ello, dado un vértice v_i , su coeficiente de *clustering* C_i se calcula como la proporción de conexiones existentes entre sus vecinos dividido por todas las conexiones posibles que se podrían dar. Esta fórmula varía ligeramente dependiendo del tipo de grafo, ya que en un grafo dirigido $G = (V, E)$, los enlaces e_{ij} y e_{ji} entre los vértices v_i y v_j no son iguales como en uno no dirigido. Por lo tanto, en el caso de un grafo dirigido, para cada vecindario N_i formado por vértices inmediatamente conectados a v_i pueden existir $\deg(v)(\deg(v) - 1)$ conexiones entre todos los vértices del vecindario, siendo $\deg(v)$ el número de vecinos de un nodo:

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{\deg(v)(\deg(v) - 1)} \quad (2.3)$$

Por otro lado, dado que en los grafos no dirigidos e_{ij} y e_{ji} son iguales, el número de conexiones posibles para un vértice v_i en este caso son $\frac{\deg(v)(\deg(v)-1)}{2}$, por lo que el coeficiente se calcula de la siguiente manera:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{\deg(v)(\deg(v) - 1)} \quad (2.4)$$

2.2. Las redes sociales como grafos

Con el fin de estudiar estas propiedades estructurales, las redes sociales se modelan como grafos aleatorios para poder realizar consideraciones estadísticas. Para ello se define una simetría entre los elementos de los grafos y los de las redes sociales, de forma que los vértices representan los actores y las aristas las relaciones entre estos. Generalmente los nodos simbolizan los usuarios de la red, aunque también pueden representar *tweets* o mensajes, mientras que las interacciones entre estos como los *retweets*, *follows* o amigos serán interpretadas por las aristas del grafo. En una red se pueden representar uno o varios tipos de relaciones entre los actores, así como relaciones unidireccionales o bidireccionales mediante el uso de grafos dirigidos y no dirigidos respectivamente.

Facebook o LinkedIn son algunos ejemplos de redes configuradas como grafos no dirigidos ya que las relaciones funcionan en ambas direcciones como “amistades”, mientras que en redes como Twitter el concepto de “seguidores” implica que las conexiones entre nodos sean dirigidas.

El interés en la estructura de estas redes reside en que estas no fueron diseñadas, sino que han ido evolucionando por sí mismas de manera aleatoria. Sin embargo, con el fin de modelar estas redes de forma precisa los grafos empleados deben tener unas características concretas.

2.2.1. Distribución *power law*

Uno de los resultados más significativos fruto del estudio de estas redes ha sido el descubrimiento de una propiedad que muestra que la distribución de la conectividad de los nodos sigue una *power law*. Esto quiere decir que mientras la gran mayoría de nodos tienen muy pocas conexiones, existe una pequeña cantidad de nodos que concentran la mayoría de estas conocidos como *hubs*.

Las distribuciones obtenidas de mediciones en Internet generalmente se pueden dividir en “cuerpo” y “cola”: el cuerpo se corresponde con los valores pequeños y medianos que representan gran parte de la distribución mientras que la cola se corresponde con el resto de valores de gran tamaño. Se dice que una distribución tiene una “cola pesada” cuando su cola no está limitada exponencialmente, esto tiene especial importancia en el modelado de la distribución ya que el cambio que supone la diferencia de ocurrencias de los valores más altos pero menos frecuentes tiene mayor influencia en esta que la existencia del resto de valores.

Por otra parte, la existencia de una “cola larga” es característica de relaciones de *power law*. Esta representa la existencia de una cantidad mucho mayor de eventos de poca frecuencia en comparación con lo observado en una distribución normal o gaussiana. Este fenómeno se puede observar en el uso de *keywords* en las búsquedas: existe un pequeño número de palabras usado con mucha frecuencia mientras que la fracción mayoritaria de las búsquedas son palabras clave usadas de forma poco común [9].

2.2.2. Propiedad de *clustering*

Una de las características más notables de las redes sociales es la existencia de *clusters*, es decir, conjuntos de individuos con una gran cantidad de conexiones entre ellos en comparación con sus conexiones con usuarios no pertenecientes al *cluster*. Estos conjuntos se forman debido a la tendencia del ser humano a formar comunidades en torno a ideas u opiniones compartidas. El interés de estos agrupamientos reside en que normalmente la presencia de una conexión entre dos nodos se debe a la existencia de algún interés común, por lo que mediante el estudio de estos *clusters* se pueden determinar estas ideas en base a los integrantes del *cluster*.

Estos grupos son de gran importancia debido a su influencia en la transmisión de información en la red, de hecho estos se emplean para el desarrollo de campañas de marketing e incluso algunos estudios afirman que su identificación puede ayudar en el reconocimiento de células terroristas [10]. Además, la propagación de noticias falsas se maximiza al unirse al comportamiento que fomentan las denominadas *echo chamber*, una forma de *cluster*. Se trata de grupos de usuarios que comparten intereses o puntos de vista y que comparten frecuentemente noticias. Estas estructuras fomentan la transmisión de aquella información que refuerza las opiniones de sus miembros, propagando este contenido mientras se censura aquel que presenta visiones contrarias.

En un estudio realizado en 2018 [11] en el que se analizaron 125 rumores de seis populares páginas de *fact-checking* con sus respectivos *tweets* y *retweets*. Esta investigación determinó que los rumores creados por miembros de echo chambers tenían una mayor probabilidad de hacerse virales y se propagaban con mayor rapidez.

En esta investigación, se emplearon dos formas distintas de caracterizar las *echo chambers*: la exposición selectiva y la homogeneidad política. La exposición selectiva es el fenómeno según el que los usuarios tienden a consumir información que quieren creer, una propiedad característica de los *echo chambers* políticos. Una vez determinada la polaridad política de cada usuario en base al contenido que publica y las cuentas a las que sigue, se observa que los miembros de *echo chambers* muestran una exposición selectiva mayor, ya sea a la izquierda o a la derecha, que aquellos que no son miembros.

En cuanto a la homogeneidad política, esta se calculó para cada miembro de una *echo chamber*, revelando que estos tienden a tener visiones políticas similares, en concreto “un 62,3 % de estos miembros manifestaban una visión política homogénea, mientras que la mitad de pares elegidos al azar apoyaban al mismo partido político”.

Para medir el papel que tienen las *echo chambers* en la propagación de rumores, se compararon cascadas de rumores en las que participaban tanto miembros como otros usuarios. Los resultados muestran que el tamaño del 91.7 % de las cascadas en las que no participan miembros es de 1, es decir, estos rumores tienen poca probabilidad de propagarse sin la intervención de miembros de echo chambers, mientras que aquellos en los que participan tienden a ser virales. También se determinó que aquellos los rumores escritos por usuarios que no participan en ninguna echo chamber son propensos a propagarse solo a otros usuarios que tampoco son miembros. Por otro lado, en el caso de rumores creados por miembros de una determinada echo chamber, estos son más propensos a difundirse entre otros miembros de esta o usuarios que no pertenecen a ninguna antes que a miembros de otra *echo chamber*.

Además, los integrantes de una echo chamber también juegan un importante papel en la velocidad con la que se propaga el rumor, ya que según esta investigación no solo retweetean con más velocidad sino que también provocan rápidas respuestas de otros usuarios.

2.3. Modelos de grafos aleatorios

En los últimos años se han propuesto una gran cantidad de modelos para replicar la estructura de las redes sociales, en la mayoría de los casos se trata de modelos que definen el comportamiento individual de los nodos con el propósito de obtener grafos con una distribución de grado como la explicada.

El primer modelo de grafo aleatorio que tuvo gran difusión fue el de Erdős-Renyi [12]. Este método considera que existe una independencia estadística en las conexiones de un nodo con el resto de la red, de forma que cada nuevo enlace se genera de manera aleatoria. Considerando una red con N nodos sin aristas entre ellos, se define un número de iteraciones M en cada cual se unen un par de nodos elegidos aleatoriamente, sin tener en cuenta aquellos que ya se encuentren enlazados. De esta forma, dependiendo del tamaño definido de M en comparación con N se obtendrán redes de características distintas. Para un valor pequeño de M resultará una red con muchos nodos desconectados y una pequeña fracción de estos formando subgrafos, mientras que en caso contrario probablemente la mayor parte de nodos se encuentren conectados entre sí. Este método genera grafos aleatorios con una distribución de grado de Poisson (en el límite de un grafo infinito). Esta es la principal limitación si queremos usar el grafo como modelo de redes sociales: la distribución de grados es muy alejada de la *power-law* que se observa en las redes sociales. En la actualidad ha servido como base para el desarrollo de otros modelos de grafos aleatorios que reproducen de manera más precisa las propiedades de estas redes.

Con objeto de utilizar estructuras cuyas características se asemejan a las de las redes sociales emplearemos los modelos desarrollados por Leskovec y Barabási para generar los grafos aleatorios con los que trabajaremos.

2.3.1. Modelo de Barabási

El estudio presentado por Albert-László Barabási y Réka Albert en [13] se enfoca en la propiedad observada en las redes consistente en la existencia de un alto grado de autoorganización que resulta en redes libres de escala. Estas redes son de un tipo concreto de red compleja cuya característica principal es que la probabilidad $P(k)$ de que un vértice del grafo esté conectado a otros k vértices sigue una ley de potencia de la forma $P(k) \sim k^{-\gamma}$, con $2 < \gamma < 3$ generalmente. La importancia de este modelo reside en que la distribución de probabilidad de vecinos de un nodo sigue el modelo *power-law*. Los modelos de grafos aleatorios de ER y de Watts y Strogatz (small world) resultan en una probabilidad $P(k)$ que decrece exponencialmente con k , es decir, en estos grafos se subestima el número de nodos que tienen un número relativamente alto de vecinos: prácticamente todos los nodos tienen muy pocos vecinos.

En investigaciones anteriores, a la hora de construir el grafo se partía de un número fijo de vértices N con conexiones creadas al azar, sin variar N , lo que no es una representación fiel del crecimiento de las redes sociales, que están en continua expansión. Además, las conexiones entre vértices se creaban de forma aleatoria y uniforme, suceso que no se observa en las redes sociales, cuyos vértices tienen una probabilidad mayor de unirse a aquellos con más aristas.

Uno de los grafos examinados en la investigación fue el de la *www* con alrededor de 800 vértices, en este los vértices representan los documentos y las aristas son los enlaces entre los documentos

existentes. La probabilidad de que k documentos apunten a una página en este grafo sigue una ley de potencia con exponente $\gamma = 2,1 \pm 0,1$.

Con el fin de replicar redes complejas como la anterior, el grafo planteado comienza con un pequeño número de vértices. En cada iteración se le añade un nuevo vértice con m_o conexiones a los vértices ya existentes. En cuanto a la probabilidad escogida para añadir aristas, esta depende del grado de conectividad $\deg(u)$ de cada vértice, siendo $\Pi(\deg(u)) = \frac{\deg(u)}{\sum_v \deg(v)}$. Esta unión preferencial es la causa de que una pequeña cantidad de nodos concentre la mayor parte de las conexiones, producido por el fenómeno “el rico se enriquece” que acentúa cualquier diferencia inicial en el grado de conectividad de dos vértices. De esta forma, los nodos más antiguos crecerán a expensas de los más jóvenes.

Al incorporar estas propiedades resulta una red libre de escala con probabilidad $P(k)$ cuyo exponente es $\gamma = 2,9 \pm 0,1$. Esta red es independiente del tiempo, aunque está en continuo crecimiento siempre se auto organiza de esta manera.

Podemos comprobar que ambas características son necesarias para desarrollar un modelo que represente de forma precisa la estructura de las redes sociales eliminando una u otra y observando el resultado. Si se parte de N vértices sin conexiones y se aplica el apego preferencial para añadir las aristas, llegará un punto en el que dado que N es constante y el número de aristas continúa creciendo, todos los nodos estarán conectados. Si por el contrario, se aplica una misma probabilidad para conectar todos los nodos, el grafo resultante perderá la propiedad de ser una red libre de escala. No obstante, este estudio no tiene en cuenta todos los aspectos de las redes, por ejemplo existen redes que no solo crecen agregando nodos sino también añadiendo y eliminando conexiones entre ellos. El modelo presentado por Leskovec presenta ciertas mejoras con respecto a este como se explicará a continuación.

2.3.2. Modelo de Leskovec

El análisis del comportamiento individual de los nodos llevado a cabo en [14] dio lugar a un modelo apoyado en anteriores basados en apego preferencial, aportando el descubrimiento de la importancia de la localidad a la hora de crear nuevas conexiones.

En esta investigación se emplearon los datasets de grandes redes sociales, de las que se contaba con información temporal sobre la creación de nuevos nodos y conexiones. Estos datos permitieron estudiar en profundidad la evolución de estos grafos conexión a conexión y calcular la probabilidad con la que se añadiría cada conexión dado un modelo en base al estado del grafo. Con el fin de comparar la precisión de cada modelo se emplea el principio de máxima verosimilitud que consiste en estudiar la evolución de un grafo paso a paso y calcular la probabilidad con la que cierto modelo hubiese producido cada conexión. De esta forma, cuánta mayor probabilidad tenga un modelo, mejor

representará la estructura de estos grafos.

Referente al apego preferencial, es decir, la tendencia con la que se eligen los nodos origen y destino de las conexiones, se consideran el grado y la edad del nodo como motivadores de esta preferencia. En cuanto al grado, a la llegada de un nuevo nodo este crea un número de conexiones constantes cuyo nodo destino es determinado de manera proporcional a su grado. Por otro lado, se teoriza que los nodos con más antigüedad tienen más experiencia en la red y por tanto crean más conexiones, por lo que también se elegirá de forma preferente aquellos nodos con mayor antigüedad como nodos destino.

Ambas propiedades fueron empleadas tanto por separado como en conjunto para dar lugar a varios modelos y poner a prueba su impacto en la generación de grafos aleatorios con el principio de máxima verosimilitud. Los resultados obtenidos para cada red social fueron distintos, aunque se obtuvieron ciertas conclusiones generalizadas. Se observó que el sesgo de grado es más débil en el nodo destino que en el nodo origen de la conexión ya que aparentemente su elección es más aleatoria. Considerando todos los modelos y sus resultados, se llegó a la conclusión de que el modelo de apego preferencial descrito en el estudio de Barabási (consistente en que la probabilidad utilizada para seleccionar un nodo como destino sigue una ley de potencia) funciona mejor que otras variantes basadas en la edad y el grado de los nodos. A pesar de que estos resultados confirman la eficacia del modelo de apego preferencial, este no considera la localidad a la hora de crear nuevas conexiones, es decir, la hipótesis de que es más probable que estas se creen entre nodos que ya tienen conexiones en común. Para medir esta localidad se examina el mínimo número de conexiones que existen entre dos nodos. Al comparar los resultados obtenidos de las redes sociales y el modelo de apego preferencial se observan claras diferencias, ya que aunque en este último prevalezcan los caminos cortos entre nodos también existe un número considerable de nodos separados por más de dos conexiones. Por otra parte, en las redes “la probabilidad de unirse a un nodo a una distancia h (h conexiones) decae exponencialmente dado que el número de aristas a esta distancia aumenta exponencialmente con h ”. La observación más notoria de estos resultados es que la mayor parte de estas conexiones se dan entre nodos que cierran triángulos, es decir, que se encuentran a dos conexiones de distancia. Este descubrimiento lleva a la cuestión de cómo seleccionar qué camino tomar cuyo resultado cierre un triángulo de entre todos los nodos existentes que cumplen estar a dos “saltos” del nodo origen.

Dado un nodo origen u y un nodo w que se encuentra a dos saltos de este, se estudiaron los resultados de varios modelos para elegir qué nodo intermedio v los conectaría y qué nodo w se seleccionaría para cerrar el triángulo. Tras probar distintos modelos para la elección de las conexiones v y w se concluyó que el modelo uniformemente aleatorio era la mejor opción en su conjunto. Su simplicidad es otra de las razones por las que se eligió este modelo frente a otros, ya que en redes no triviales el rendimiento decae a medida que se calculan las probabilidades de nodos cada vez más lejanos al aumentar drásticamente el número de posibles caminos.

Una vez definida la metodología para crear conexiones entre nodos, en este estudio se presenta cómo determinar qué nodos serán responsables de crear estas conexiones. Lo primero que se precisa es el tiempo de vida de un nodo, que sigue una distribución exponencial $pl(a) = \lambda \exp(-\lambda a)$ siendo el tiempo de vida de un nodo u $a(u)$. Aunque esta distribución no se ajuste de manera correcta a nodos con poca longevidad en la red, sí representa con precisión nodos cuya edad sea mayor a una semana. A la hora de calcular los intervalos de tiempo entre la creación de conexiones por un nodo se determinó que la mejor probabilidad se obtiene con una ley de potencia con un corte exponencial $p_g(\delta(d); \alpha, \beta) \propto \delta(d)^{-\alpha} - \alpha^{(-\beta\delta(d))}$ siendo d el grado del nodo. Cabe destacar que los intervalos de tiempo entre la creación de conexiones por un nodo tiene una relación directa con su antigüedad en la red, ya que nodos más antiguos tienden a crear conexiones con intervalos de inactividad mucho menores. Finalmente, en cuanto al tiempo de llegada de cada nodo, en esta investigación se llega a la conclusión de que este puede ser desde sublineal hasta exponencial dependiendo de la red social, por lo que deberá ser especificado de antemano.

2.4. Modelos de difusión

2.4.1. Modelo epidemiológico

Derivado del estudio de propagación de infecciones, se han desarrollado distintos modelos epidemiológicos basados en ecuaciones diferenciales. Algunos de los modelos más básicos son Susceptible Infected (SI) y Susceptible Infected Susceptible (SIS) dependiendo de la enfermedad a estudiar [15]. En estos se considera una enfermedad contagiosa por lo que una vez infectado un individuo susceptible este es infeccioso inmediatamente, teniendo todos los individuos la misma probabilidad de infectarse. Esta enfermedad no confiere inmunidad una vez recuperados por lo que dependiendo del modelo se mantendrán infecciosos en el tiempo o se recuperarán y volverán a ser susceptibles a la infección.

En ambos modelos existen dos estados: susceptible de infectarse al no tener inmunidad frente a la enfermedad e infectado y por tanto infeccioso. En estos se define la probabilidad de que un individuo infectado contagie a uno susceptible mediante la tasa de infección β . Si los individuos se mueven libremente, el producto $S(t)I(t)$ es proporcional a la probabilidad de que un susceptible encuentre un infectado. Por otro lado, la probabilidad de recuperación γ se emplea únicamente en el modelo SIS ya que es la que permite a un individuo volver a ser susceptible tras haberse recuperado de la enfermedad. Esta se define como $\gamma = \frac{1}{D}$ siendo D el periodo de duración de la infección.

A continuación se va a explicar la versión más simple de los modelos que no considera la duración del periodo vital de los individuos desde su nacimiento hasta su muerte.

En el caso del modelo SI o Susceptible-Infected los individuos nunca generan inmunidad frente a

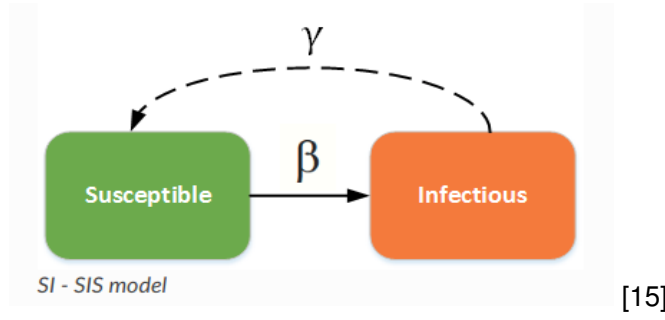


Figura 2.1: Estados de los modelos SI-SIS [15]

la enfermedad por lo que una vez el nodo se haya infectado, este se mantendrá en este estado a lo largo del tiempo, manteniéndose en contacto con el resto de la población y contribuyendo al proceso de propagación.

Siendo S el número de individuos susceptibles e I el número de infectados, la población total se considera $N = S + I$ por lo que el modelo SI se puede expresar como:

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta SI}{N} \\ \frac{dI}{dt} &= \frac{\beta SI}{N} = \beta I \left(1 - \frac{I}{N}\right)\end{aligned}\quad (2.5)$$

El modelo SIS o Susceptible-Infected-Susceptible sin embargo, tiene en cuenta infecciones de las que un individuo se puede recuperar sin generar inmunidad y por tanto cuyos afectados son siempre susceptibles de infectarse de nuevo. Por lo tanto en este caso existen los estados susceptible e infectado, con las mismas implicaciones que en el modelo anterior bajo la diferencia de que un individuo infectado puede volver al estado susceptible.

$$\begin{aligned}\frac{dS}{dt} &= \mu N - \frac{\beta SI}{N} - \gamma S \\ \frac{dI}{dt} &= \beta I \left(1 - \frac{I}{N}\right) - \gamma I\end{aligned}\quad (2.6)$$

En este, para que la enfermedad continúe propagándose en la población es necesario que $\frac{\gamma}{\beta} > 1$, ya que en caso contrario esta termina por desaparecer. Por lo tanto, existen dos estados de equilibrio en este modelo: una población libre de infectados donde $I = 0$ y un estado estable cuando $\frac{\partial I}{\partial t} > 0$ como se muestra en 2.7:

$$I = \frac{(\beta - \gamma)N}{\beta} = \left(1 - \frac{\gamma}{\beta}\right)N \quad (2.7)$$

2.4.2. Modelo de difusión de desinformación en las redes sociales

Dada la contribución de las redes sociales a la aceleración del proceso de difusión de información y en particular de desinformación y rumores, el objetivo del modelo [16] es estudiar cómo la existencia de mecanismos de verificación de la información pueden contener su propagación. En este modelo nos alejamos de la epidemiología ya que este carece de sentido desde el punto de vista de un virus, este se trata de un modelo específico para el estudio de la información. Para ello se han definido tres estados posibles para cada nodo del grafo: susceptible, creyente o *believer* y no creyente o *fact-checker*.

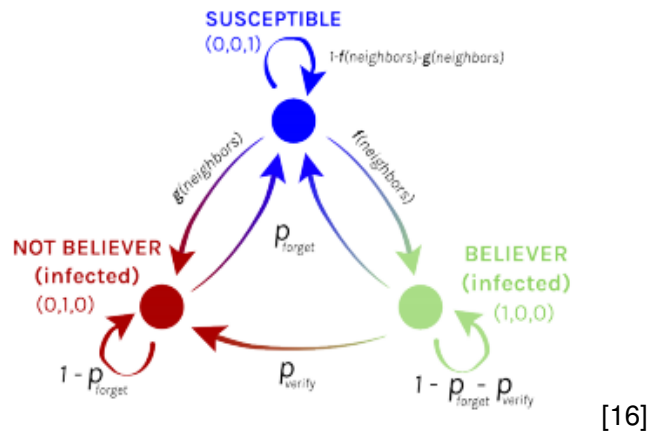


Figura 2.2: Estados del modelo de difusión de desinformación

Un nodo susceptible puede convertirse en creyente o no creyente, o mantenerse en el estado actual, en función de los estados en los que se encuentren sus vecinos, mediante las respectivas funciones de difusión f_i y g_i . En estas, se emplean la variable fijas $\beta \in [0, 1]$ que indica la tasa de difusión y $\alpha \in [0, 1)$ que expresa la credibilidad del rumor, así como n_i^B y n_i^F que son el número de vecinos de i que son *believers* o *fact-checkers* en t respectivamente.

Una vez un nodo pasa al estado *fact-checker*, este solo puede mantenerse en este estado u olvidar la información y volver a ser susceptible a esta. Para calcularlo se emplea la variable fija p_{forget} que define la probabilidad con la que un nodo deja de tener una opinión.

Finalmente, un nodo creyente puede pasar a todos los estados posibles definiendo además la variable constante p_{verify} que determina la probabilidad de que este averigüe que la información es falsa y por lo tanto se convierta en *fact-checker*, empleando la probabilidad p_{forget} mencionada según la cual el nodo volvería a ser susceptible.

$$f_i(t) = \beta \frac{n_i^B(t) \cdot (1 + \alpha)}{n_i^B(t) \cdot (1 + \alpha) + n_i^F(t) \cdot (1 - \alpha)} \quad (2.8)$$

$$g_i(t) = \beta \frac{n_i^F(t) \cdot (1 + \alpha)}{n_i^B(t) \cdot (1 + \alpha) + n_i^F(t) \cdot (1 - \alpha)} \quad (2.9)$$

DISEÑO Y DESARROLLO

3.1. Velocidad de difusión

La velocidad de difusión de información consiste en una medición propuesta en este trabajo con el fin de determinar cómo afectan ciertas características estructurales de los grafos aleatorios al tiempo necesario para alcanzar un número definido de nodos N conocedores de la información.

La información será considerada como una opinión consistente en creer o no creer que la información es correcta, por lo que el nodo se podrá encontrar en distintos estados dependiendo de su comportamiento frente a esta. Para la formación de opiniones se han empleado distintos modelos: desde los modelos SI y SIS hasta el modelo de difusión de desinformación.

En cuanto al tiempo de propagación este no se refiere a tiempo real sino al número de iteraciones que se llevan a cabo sobre el grafo, es decir, la cantidad de veces que se es necesario que se recorra hasta alcanzar el número de nodos infectados definido.

Se trata de un proceso estocástico a tiempo discreto, es decir, un conjunto de variables aleatorias con su propia distribución de probabilidad que son dependientes de un parámetro, en este caso el tiempo. Este proceso es característico de series temporales como la evolución de la población en un territorio año tras año.

En concreto, este modelo presenta la propiedad de Markov [17]. Los procesos que presentan esta propiedad son comúnmente conocidos como procesos "sin memoria", ya que en estos la probabilidad de que una variable cambie de estado en el futuro depende únicamente del valor que esta tiene en la actualidad, sin tener en cuenta los valores que ha podido tomar históricamente [18].

De esta manera, este se trata de un modelo o cadena de Markov [19] $X = (X_n)_{n \in N} = (X_0, X_1, X_2, \dots)$ tal que en cada instante el proceso toma sus valores en un conjunto discreto M siendo $X_n \in M$ y $\forall n \in N$. De esta manera, la propiedad de Markov implica que:

$$P(X_{n+1} = S_{n+1} \mid X_n = S_n, X_{n-1} = S_{n-1}, X_{n-2} = S_{n-2}) = P(X_{n+1} = S_{n+1} \mid X_n = S_n) \quad (3.1)$$

Una vez definidos estos conceptos, el proceso de difusión tiene lugar definiendo dos estados posibles para un nodo: sin informar e informado, que consideraremos estados 0 y 1 respectivamente. Una vez definidos estos conceptos, el flujo de trabajo del modelo es el siguiente.

Partiendo de un grafo aleatorio cuyos nodos se encuentran todos en estado 0, indicativo de que ninguno porta la información, se elige un nodo de este grafo de forma totalmente aleatoria al que asignar el estado 1 o, de forma análoga, al que "infectar" con la opinión. Este nodo debe tener vecinos a los que poder propagar la información, en caso contrario se elegiría otro nodo al que infectar. A continuación se comienza a iterar sobre la lista de nodos pertenecientes al grafo de forma que en cada iteración se obtienen todos los vecinos del nodo en cuestión y se computa el estado de este en base al modelo de formación de opiniones escogido.

Una vez se haya recorrido el grafo en su totalidad se incrementa en uno el tiempo de propagación y se comprueba el número de nodos en estado 1 o portadores, finalizando al alcanzar la cifra de infección definida.

3.2. Formación de opiniones

3.2.1. Modelo SI

En el caso del modelo SI, se define una variable β fija antes de comenzar el proceso de propagación cuyo valor determinará la probabilidad con la que el nodo toma la información. Esta probabilidad se calcula si al menos uno de los vecinos del nodo en cuestión está "infectado".

Siendo n_i el número de vecinos del nodo que se encuentran en estado 1, la probabilidad de difusión es:

$$P_d = 1 - (1 - \beta)^{n_i} \quad (3.2)$$

De esta forma, si el nodo no está infectado, es decir, se encuentra en estado 0, se genera una probabilidad aleatoria y si esta es menor que P_d el nodo pasará al estado 1 mientras que en caso contrario se mantendrá en su estado original. Sin embargo, al tratarse del modelo SI, en caso de que el nodo en cuestión ya se encuentre en estado 1 o infectado este ya no cambiará, únicamente contribuyendo a la infección de sus vecinos.

3.2.2. Modelo SIS

En este modelo se define la variable β y se calcula la probabilidad de difusión P_d al igual que en el modelo SI para los nodos en estado 0.

No obstante, se opera de forma distinta en lo referente a los nodos ya infectados: para ello se define también la probabilidad δ . Este parámetro indica la probabilidad de que un nodo olvide la información y vuelva al estado original 0. De esta manera se obtiene la variable τ como la razón entre β y δ , que tendrá un valor entre 1 y 2 ya que la probabilidad de olvidar la información siempre será inferior a la probabilidad de que esta se difunda:

$$\tau = \frac{\beta}{\delta} \quad (3.3)$$

Para ello se genera una probabilidad aleatoria y si esta es menor que δ el nodo se "recupera", en caso contrario el nodo mantiene su opinión original.

3.2.3. Modelo de difusión de desinformación

Para implementar el modelo de difusión de desinformación es necesario introducir un nuevo estado que represente la posibilidad de que un individuo averigüe que la información es falsa y actúe por lo tanto como *fact-checker*.

En este modelo además de definir las variables fijas β y τ , se define la probabilidad p_{verify} de que el nodo verifique la información. Por lo tanto, se pueden dar tres comportamientos posibles frente a la información: su difusión, su verificación o su omisión.

La difusión se da si un nodo se encuentra en estado 0 y pasa a conocer la información y creerla (*believer*) o a verificar que esta es falsa (*fact-checker*). Esto sucede mediante las probabilidades P_b y P_f respectivamente siendo n_b el número de vecinos que ya son *believers* y n_f los que son *fact-checkers*.

$$\begin{aligned} P_b &= 1 - (1 - \beta)^{n_b} \\ P_f &= 1 - (1 - \beta)^{n_f} \end{aligned} \quad (3.4)$$

Si la probabilidad aleatoria calculada es menor que P_b , el nodo pasará a creer la información, en otro caso si es menor que P_f este sabrá que es falsa y por último en caso de que ninguno de los casos anteriores se cumpla el nodo mantendrá su estado original.

Por otra parte, la verificación sucede en el caso de que un individuo *believer* pase a ser *fact-checker*

mediante la probabilidad fija p_{verify} , mientras que si la probabilidad aleatoria obtenida es menor que δ el nodo olvidará la información. De la misma manera, un individuo que conoce la falsedad de la información puede olvidarla también mediante la probabilidad δ .

3.3. Medidas

3.3.1. Medidas según β

Se ha estudiado la velocidad de difusión según la variable fija β cuyo valor indica la probabilidad de infección de un nodo. Para ello se han definido dos grafos aleatorios siguiendo los modelos de Barabási y Leskovec respectivamente y se ha observado el efecto de variar los valores de β a 0'01, 0'03, 0'05 y 0'07 en cada grafo.

3.3.2. Medidas según β y δ

Con el fin de estudiar los efectos de un modelo ligeramente más complejo y realista como es el SIS se ha incluido la variable fija τ para calcular la probabilidad de que el nodo olvide la información como $\delta = \frac{\beta}{\tau}$. Para ello se ha fijado $\beta = 0'02$ ya que este se ha considerado un valor intermedio y se ha variado τ a 1'0, 1'2, 1'4 y 1'6.

3.4. Modificaciones de grafos

Los modelos SIS y SIR en espacios continuos asumen una completa libertad de movimiento, por tanto dependen sólo de los parámetros epidemiológicos β y δ (o, según el modelo, τ). Sin embargo, en un grafo la posibilidad de contacto entre dos individuos está determinada por la estructura del grafo mismo que por tanto juega un papel fundamental en la difusión. Existen varios estudios al respecto con resultados parciales en este campo. En un estudio sobre la propagación de virus [20] se demuestra que, para el modelo SIS, el umbral de difusión de un virus es $\tau > 1/\lambda_1$, donde λ_1 es el autovalor más grande o radio espectral de la matriz de adyacencia del grafo. De hecho, se han obtenido resultados análogos [16] para la difusión de información con *fact-checking*.

Por lo tanto, dado que se ha evidenciado que al cambiar el radio espectral también lo hace el umbral de difusión, es de especial interés estudiar cómo la estructura de un grafo incide en la difusión. Con objeto de medir qué propiedades estructurales tienen una mayor repercusión sobre la velocidad de difusión, se buscan en concreto alteraciones que modifiquen estos autovalores sustancialmente. De forma general, la eliminación de nodos del grafo, o de forma análoga, el *fact-checking* de rumores por

el que un nodo verifica un rumor y no lo retransmite, causan que los autovalores de la matriz de adyacencia del grafo disminuyan y por lo tanto también lo haga la velocidad de difusión. En consecuencia, dado que se quieren eliminar k nodos, una solución óptima sería identificar un conjunto de nodos que ralentice lo máximo posible la propagación.

Aunque no existen resultados específicos en este campo, en un estudio similar [7] se estudió la propagación de campañas competidoras en redes sociales, donde una de las campañas difunde desinformación. En este caso se buscaba identificar el conjunto de individuos a los que forzar a adoptar la campaña verdadera que serían necesarios para limitar la difusión de la otra campaña. Sin embargo este problema resultó NP-completo y por tanto, imposible en la práctica, por lo que se definieron tanto un algoritmo codicioso como heurísticas poco costosas con el fin de obtener soluciones aproximadas. Dado que el método codicioso no es escalable a grandes redes sociales debido a su coste, se comparó su rendimiento con el de estas heurísticas, consiguiendo resultados similares. Entre estas, la heurística con mayor éxito fue aquella que elegía los nodos de mayor grado de la red como los nodos más influyentes del grafo para la propagación de la campaña deseada.

Por consiguiente, en este estudio se va a analizar el efecto de soluciones heurísticas como son el eliminar los nodos con mayor índice de una medida concreta en el grafo. En especial, los nodos con más vecinos y mayor coeficiente de *clustering* dada la distribución *power law* de las redes planteadas. Asimismo, como control y con el fin de contrastar y verificar la validez de los resultados obtenidos, también se van a eliminar los nodos con menor índice de las medidas especificadas.

Habiendo definido las variables fijas β y τ y empleando por tanto un modelo equivalente al modelo epidemiológico SIS se ha decidido modificar la estructura de los grafos empleados con el fin de estudiar el impacto de propiedades como la vecindad y el *clustering* sobre la velocidad de propagación en grafos generados por el modelo de Barabási y el de Leskovec.

Para ello, se ha decidido eliminar progresivamente los nodos con mayor y menor vecindad hasta alcanzar el 8 % del número total de nodos del grafo en cada caso, aplicando el mismo criterio para el *clustering*. Se ha fijado como tope el 8 % ya que se ha observado que generalmente una vez superada esta cifra el grafo comienza a dividirse en subgrafos no conectados entre sí, por lo que no existe forma de que la información se propague si esta no puede alcanzar al resto de nodos al no hallar un "puente" que los conecte.

PRUEBAS Y RESULTADOS

4.1. Características de los grafos generados

Para el estudio de la velocidad de difusión se han creado grafos aleatorios según los modelos de Barabási y Leskovec con las propiedades que se explican a continuación.

Los grafos generados presentan las características propias de las redes libres de escala, siguiendo una distribución de grado en forma de *power law*. Estas redes se caracterizan por la presencia de *hubs* de gran tamaño, es decir, de un pequeño conjunto de nodos altamente conectado con el resto de nodos del grafo.

Con el fin de comprobar estas propiedades se ha representado la distribución del grado de ambos grafos, calculando la probabilidad $P(k)$ de que un nodo del grafo tenga k vecinos. Para ello se han generado dos grafos siguiendo los modelos de Barabási y Leskovec respectivamente, de 5000 nodos cada uno.

En el grafo generado según el modelo de Barabási se puede observar en la figura 4.1(a) cómo el grado más común entre los nodos es 7, representando el 20 % de estos. Si sumamos las probabilidades de que un nodo tenga grado 10 o inferior esto representa el 56,78 % del grafo, con la probabilidad restante correspondiendo a ocurrencias de grados con probabilidades de entre el 0 % y el 6 %, siendo el grado máximo de un nodo alrededor de 200.

En cambio, en el grafo creado según el modelo de Leskovec mostrado en la figura 4.1(b) existen dos grados con probabilidades muy similares: el 2 y el 3 con $P(2) = 0,1006$ y $P(3) = 0,0988$, lo que implica que alrededor del 20 % de los nodos del grafo tienen un grado comprendido entre 2 y 3. Si en este caso se agregan las probabilidades de tener grado 10 o inferior esto representa el 69,54 % del grafo, siendo las probabilidades restantes entre el 0 % y el 2 % para cada ocurrencia de grado, con un grado máximo superior a 800.

En ambos casos se puede observar que la distribución tiene una "cola larga" debido a la presencia de nodos con un grado mucho mayor al resto. La existencia de un pequeño conjunto de nodos con grados alrededor de 200 u 800 respectivamente frente a la concentración de grados inferiores a 10 por

parte del resto de los nodos son una de las principales propiedades de las redes libres de escala. Estas se caracterizan por la presencia de estas diferencias en orden de magnitud. En redes no dirigidas como estas se puede representar la distribución de grado como: $P_{deg}(k) \propto k^{-\gamma}$ siendo γ un exponente. La probabilidad de cada grado $P_{deg}(k)$ disminuye lentamente a medida que k aumenta, lo que implica una menor probabilidad de encontrar nodos con un grado muy alto.

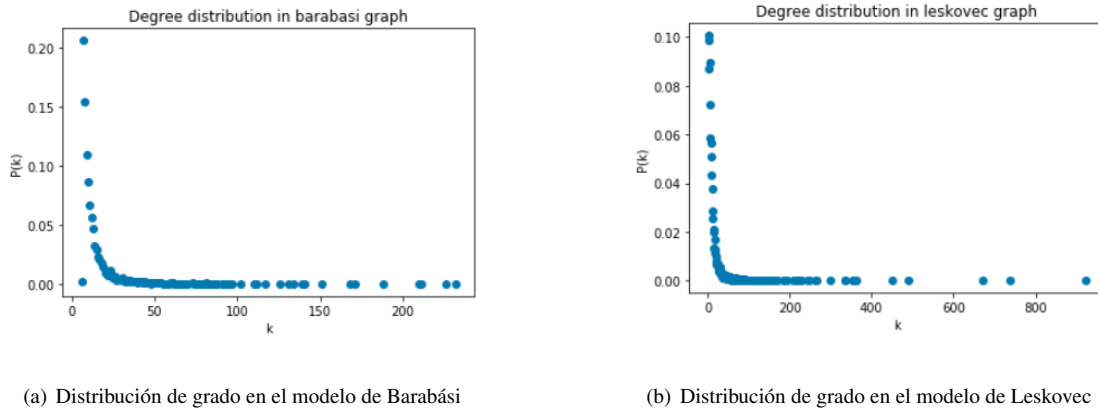


Figura 4.1: Distribuciones de grado.

Las distribuciones que siguen una *power law* son fácilmente reconocibles al emplear la escala logarítmica para representar ambos ejes. De esta forma se puede observar que la distribución de grado sigue una *power law*, ya que los puntos se agrupan formando una especie de línea, lo que significa que las variables $P_{deg}(k)$ y k muestran una relación lineal entre ellas. Aunque como se puede comprobar en las figuras 4.2(a) y 4.2(b), el diagrama pierde claridad en su forma al acercarse a los grados mayores, ya que en estos casos existen menos puntos que puedan contrarrestar el ruido.

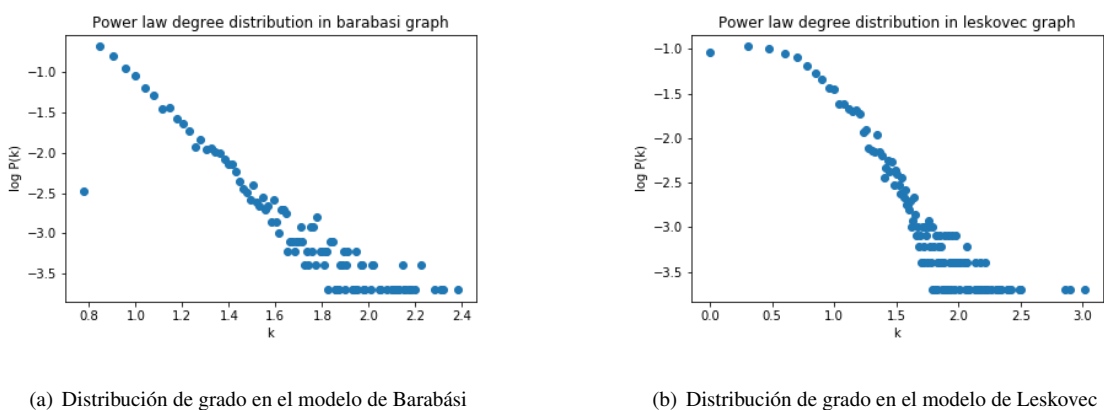
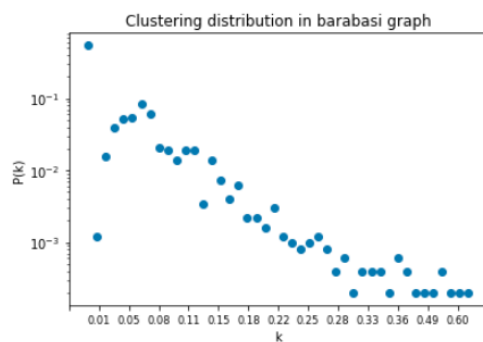
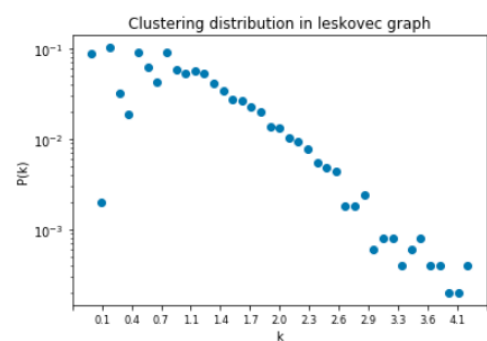


Figura 4.2: Distribuciones de grado tomando ejes logarítmicos.

Para medir el *clustering* de los grafos se ha computado el coeficiente de *clustering* que determina la proporción de los vecinos de un nodo que se encuentran conectados, para ello se ha empleado la intersección entre las listas de adyacencia de cada nodo. La distribución del *clustering* a lo largo de

la red permite observar la existencia de zonas con mayor y menor *cluster* en el grafo, siendo $P(x)$ la probabilidad de que un nodo tenga coeficiente de *cluster* x . Con el fin de medir las ocurrencias de cada coeficiente estos se han aproximado. En el grafo generado por el modelo de Barabási ha sido suficiente con una aproximación de dos decimales ya que al presentar un 43 % de nodos con un grado superior a 10 en general el tamaño de sus listas de adyacencia es mayor y por consiguiente también lo es el denominador del coeficiente. Sin embargo para calcular la distribución en el grafo de Leskovec se ha utilizado solo un decimal para el redondeo ya que este presenta casi un 70 % de nodos con grados inferiores a 10, lo que implica que el denominador del coeficiente es menor.

La distribución de *clustering* en estos grafos muestra una distribución *power law* al igual que la distribución de grado, ya que existe un pequeño número de nodos con un coeficiente de *clustering* mucho mayor que el resto, mientras la mayoría tienen un coeficiente de *clustering* menor. En el grafo de Barabási mostrado en la figura 4.3(a) se puede observar que el 91 % de los nodos tienen un coeficiente de *clustering* de 0,1 o menor, mientras que el 9 % restante abarcan coeficientes de 0,1 a 0,67 con probabilidades entre 0 % y 1,6 %. Sin embargo, en el segundo grafo creado según el modelo de Leskovec en la figura 4.3(b) los coeficientes de *clustering* son considerablemente más altos en general, con un 66 % de los nodos presentando un coeficiente entre 0 y 1, mientras que el resto comprenden coeficientes de 1 a 4 con probabilidades entre 0 % y 5 %. Esta diferencia entre los coeficientes de ambos grafos se debe al modelo empleado para generarlos. El modelo de Barabási se caracteriza por emplear *preferential attachment* para generar las conexiones entre sus nodos mientras que el modelo de Leskovec también tiene en cuenta la localidad de las conexiones, es decir, está orientado a cerrar triángulos, por lo que sus nodos presentarán un *clustering* mayor.

(a) Distribución de *clustering* en el modelo de Barabási(b) Distribución de *clustering* en el modelo de Leskovec**Figura 4.3:** Distribuciones de *clustering*.

4.2. Análisis cualitativo de la difusión

Al partir de un grafo finito cuyos nodos son susceptibles en su totalidad y cuya infección sigue una probabilidad, se puede esperar que la difusión de información siga una distribución exponencial al comienzo, con parámetro $p > 0$ el número medio de vecinos por nodo. Sin embargo, a medida que la información se difunde y más nodos son infectados, el número de nodos susceptibles a la información disminuye, lo que provoca que la expansión se ralentice hasta finalizar cuando todos los nodos hayan sido infectados. Al igual que en el modelo SI se supondrán dos estados: nodos infectados y nodos susceptibles, considerando que un nodo infectado se mantiene de esta manera sin opción a cambio. En este caso, esto se puede aplicar también al modelo SIS, ya que aunque se introduzca la existencia de la probabilidad de olvidar la información, si esta se propaga a todos los nodos del grafo la única diferencia es la velocidad con la que lo hace, ya que ralentiza el proceso de difusión.

Siendo $I(t)$ el número de nodos infectado en el tiempo t , haremos una aproximación considerando una variable continua en el tiempo. Dado un tiempo t , un grafo de N nodos contará con un conjunto de nodos infectados I , compuestos por $I(t)$ nodos, y un conjunto de nodos susceptibles S siendo $|S| = N - I(t)$. Cada nodo del conjunto I cuenta con un número medio de vecinos ρ . Siendo la probabilidad de contagio β , cada uno infectará a ρ de estos nodos, que por consiguiente tendrán $\beta\rho I(t)$ vecinos. Sin embargo, muchos de estos nodos ya estarán infectados por lo que no contribuirán al proceso de difusión.

Teniendo un conjunto de N elementos y un subconjunto M de m elementos, si se escoge un conjunto Q de q elementos al azar, el número medio de elementos de Q que pertenecen a M es $\frac{qm}{N}$. Si se aplica lo anterior en este caso, el conjunto Q se puede identificar como el conjunto de vecinos con $q = \beta\rho I(t)$ y M como el conjunto de nodos susceptibles siendo $m = N - I(t)$. Por lo tanto el número medio de nuevos nodos infectados sería:

$$\beta\rho I(t)(N - I(t)) \quad (4.1)$$

Esto proporciona la variación instantánea en el número de nodos, por lo que se puede expresar:

$$\frac{\partial I}{\partial t}(t) = \beta\rho I(t)(N - I(t)) \quad (4.2)$$

Dado que se quiere obtener la fracción del grafo que está infectada se define la cantidad $i(t) = \frac{I(t)}{N}$ y su respectiva ecuación:

$$i'(t) = \frac{\partial i}{\partial t}(t) = \frac{\beta\rho}{N} i(t)(1 - i(t)) \quad (4.3)$$

Partiendo de $t = 0$, donde solo un nodo se encuentra infectado, es decir $i(0) = \frac{1}{N}$ y suponiendo que en t infinito todo el grafo lo está, se obtiene:

$$\left\{ \begin{array}{l} \frac{\partial i}{\partial t}(t) = \frac{\beta\rho}{N}i(t)(1-i(t)) \\ i(0) = \frac{1}{N} \\ \lim_{t \rightarrow \infty} i(t) = 1 \end{array} \right\} \quad (4.4)$$

Cuya solución general es: $i(t) = \frac{1}{1+\exp(-\alpha(t-t_0))}$ siendo $\alpha = \frac{\beta\rho}{N}$.

Empleando la condición en $i(0)$

$$\frac{1}{1+\exp(\alpha t_0)} = \frac{1}{N} \quad (4.5)$$

se obtiene t_0 tal que:

$$t_0 = \frac{1}{\alpha} \log(N-1) = \frac{N}{\beta\rho} \log(N-1) \quad (4.6)$$

En $t = 0$, $i'(t) = \frac{\beta\rho}{N} \frac{1}{N} \frac{N-1}{N} = \frac{\beta\rho}{N^2}$, lo que quiere decir que la derivada es muy pequeña y crece de manera muy rápida, aproximándose a una exponencial.

Para $t = t_0$, $i(t) = \frac{1}{2}$, $i'(t) = \frac{\beta\rho}{4N}$ alcanza su valor más alto, a partir del cual la derivada disminuye y ya no presenta un crecimiento aproximado al exponencial.

Con el fin de eliminar estos efectos y poder estudiar el comportamiento exponencial de la función se puede limitar su área tal que $i(t) < \frac{1}{3}$, es decir, limitar la función hasta alcanzar entre un tercio y la mitad de nodos infectados en el grafo. Esto limita la función al área en la que presenta un crecimiento exponencial, eliminando aquella en la que se ha alcanzado tal número de nodos infectados que su crecimiento decae rápidamente.

Este fenómeno se puede observar en la figura 4.8(a), que muestra el número de iteraciones necesarias en un grafo de 5000 nodos para alcanzar la infección de la mitad de los nodos variando la probabilidad de infección entre 0,01 y 0,07. En esta se aprecia un comportamiento exponencial hasta que alrededor de 2000 nodos del grafo han sido infectados, lo que corresponde con el 40 % de los nodos del grafo. Una vez se supera este límite se puede percibir cómo este crecimiento se detiene ya que la velocidad de difusión disminuye sustancialmente. Esto se debe a la saturación del grafo antes mencionada, que provoca que una vez se haya alcanzado un cierto número de nodos infectados, disminuya la velocidad de propagación debido a la dificultad de encontrar nodos susceptibles a la infección. Una

vez el grafo se ha saturado, la probabilidad de infección no impacta de la manera en la que lo haría al comienzo de la difusión cuando existían suficientes nodos para continuar la propagación de manera exponencial. Es por este motivo que al comparar el número de iteraciones que han sido necesarias para alcanzar el 50 % de infección en el grafo con cada probabilidad no se puedan obtener resultados concluyentes como se puede observar en la gráfica 4.4(b), ya que en esta situación dar con *clusters* de nodos susceptibles tiene mayor impacto que el tamaño de la probabilidad definida, que no tiene valor si no existen nodos alrededor a los que infectar.

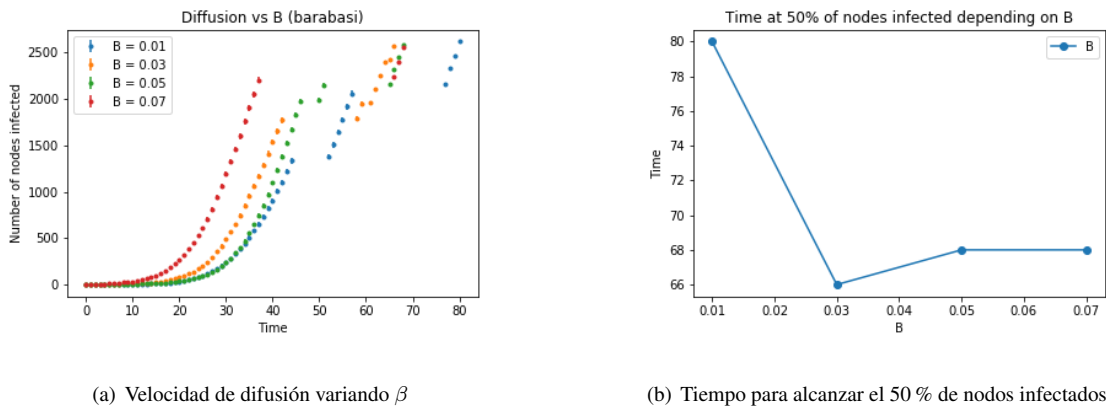


Figura 4.4: Velocidad de difusión según el modelo SI

Es por esto que en el estudio realizado a continuación se va a recortar la gráfica al 30 % o 40 % de infección en caso de que la saturación del grafo impacte los resultados obtenidos. Como se muestra en la figura 4.6(a), al recortar la gráfica anterior al 40 % se limita la función al área en la que presenta un crecimiento exponencial, lo que resulta en que la gráfica 4.6(b) muestre un decrecimiento casi lineal al aumentar la probabilidad de infección. Esto es de esperar ya que el número de iteraciones necesarias para alcanzar la cifra de nodos infectados definida disminuye al aumentar la probabilidad con la que un nodo se puede infectar.

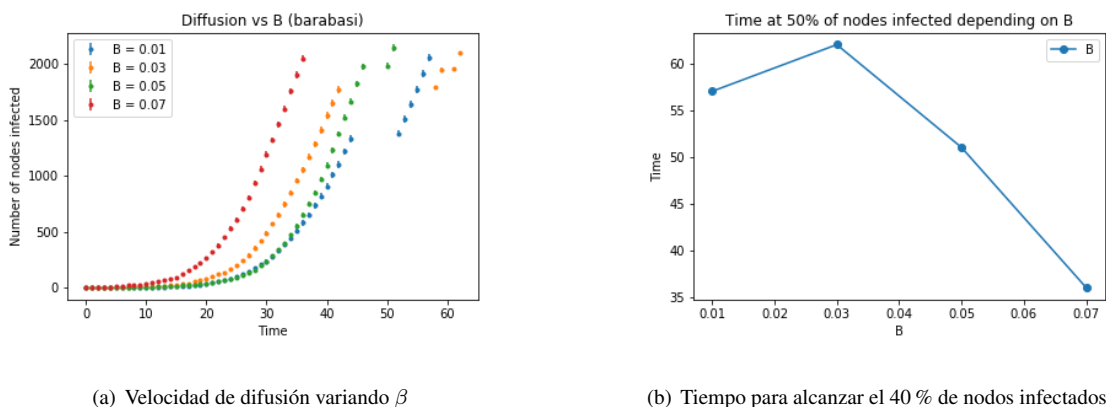
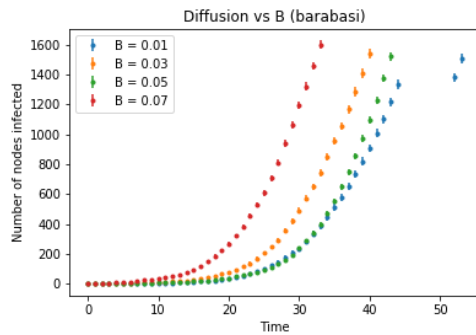


Figura 4.5: Velocidad de difusión según el modelo SI acotada al 40 % de los nodos

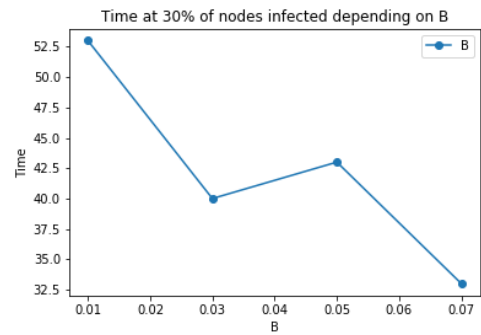
4.3. Velocidad de difusión siguiendo el modelo epidemiológico SI

En las gráficas 4.6(a) y 4.7(a) se puede observar el crecimiento exponencial de la propagación al emplear la probabilidad de infección β sin posibilidad a que los nodos olviden la información.

Para representar las gráficas 4.6(b) y 4.7(b) se han medido los valores del tiempo que se ha necesitado para que la información alcance el 30 % de la población para cada valor de β en los modelos de Barabási y Leskovec. En ambas se advierte una relación casi lineal entre la probabilidad de infección y el tiempo empleados. Esto es de esperar ya que al aumentar β aumenta la probabilidad de que un nodo se contagie con cada contacto con la información. Sin embargo, este no es el único factor a tener en cuenta, ya que la infección de un nodo no depende únicamente de β sino que está determinada también por el número de vecinos que se encuentren contagiados, es por ello que debido a la estructura irregular de los grafos aleatorios nunca se va a presentar una relación perfectamente lineal.

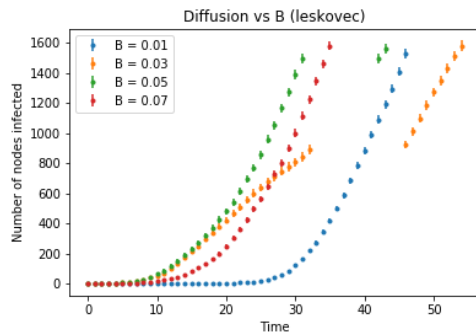


(a) Velocidad de difusión variando β

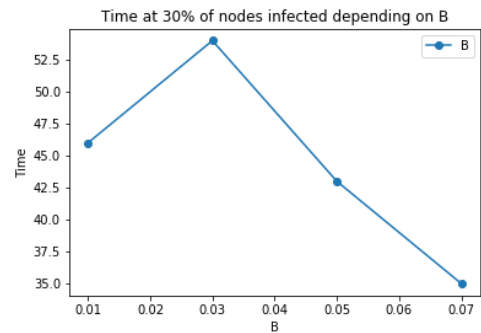


(b) Tiempo para alcanzar el 30 % de nodos infectados

Figura 4.6: Velocidad de difusión según el modelo SI en Barabási



(a) Velocidad de difusión variando β



(b) Tiempo para alcanzar el 30 % de nodos infectados

Figura 4.7: Velocidad de difusión según el modelo SI en Leskovec

La elección de un modelo u otro para la generación del grafo en este periodo del contagio no influye apenas en el tiempo que ha tomado alcanzar el 30 % de los nodos. Esto se debe a que al inicio de la propagación no existen apenas nodos infectados por lo que las diferencias estructurales entre ambos repercuten de forma mínima en la transmisión.

Sin embargo, tras superar aproximadamente el 30 % del grafo infectado se pueden observar diferencias entre las figuras 4.8(a) y 4.8(b) que comienzan a mostrar cómo el crecimiento se ralentiza, en especial en el modelo de Leskovec. Este emplea considerablemente más tiempo que el grafo del modelo de Barabási, lo que tiene su justificación en la distribución de grado en sus nodos. Como se ha expuesto al estudiar las características estructurales de ambos grafos, el modelo de Leskovec cuenta con un 20 % de nodos con grados 2 y 3, mientras que en el modelo de Barabási el grado más prevalente es el 7, que igualmente representa el 20 % del grafo. La predominancia de nodos con mayor grado en el modelo de Barabási provoca el aumento de la probabilidad de contagio de cada nodo, que depende del número de vecinos infectados.

Asimismo, también es perceptible la existencia de periodos más largos en los que el número de infectados no aumenta en el grafo generado según el modelo de Leskovec. Esto puede explicarse gracias a la diferencia entre la distribución de *clustering* de ambos grafos: en el modelo de Barabási el coeficiente más común, representando un 91 % del grafo, abarca de 0 a 0,1 mientras que en el de Leskovec este comprende de 0 a 1, representando el 66 %. Por lo tanto, la presencia de estas etapas se puede atribuir a la existencia de *clusters* mucho mayores en el segundo modelo. Esto causa que el crecimiento sea notablemente más lento cuando la "infección" o información sale de un *cluster* y entra en otro en el que aún no existen contagios.

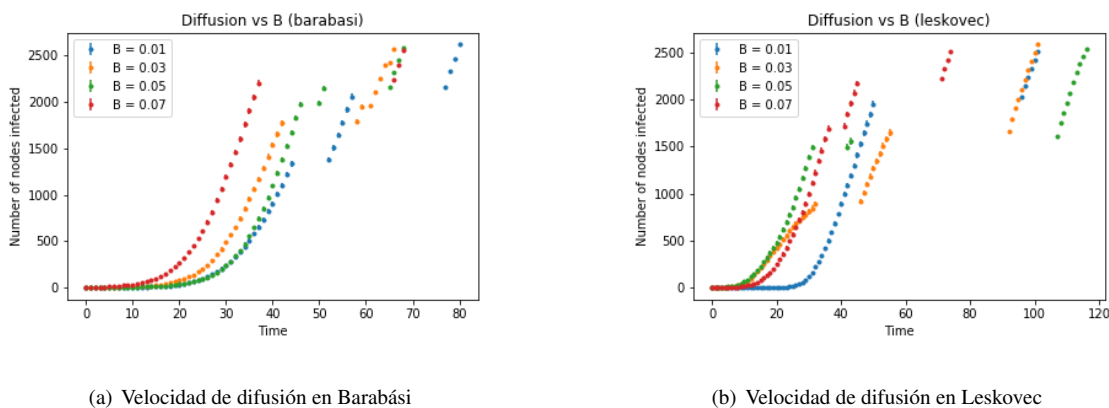
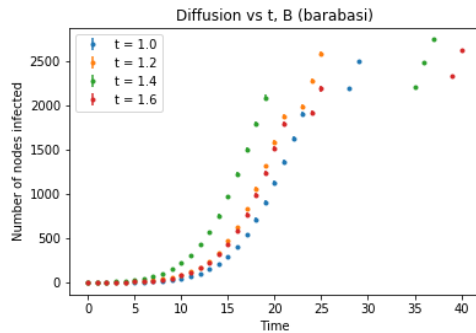


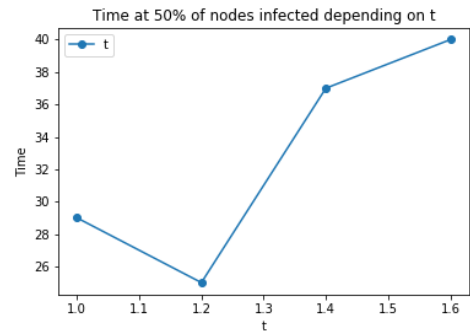
Figura 4.8: Velocidad de difusión según el modelo SI hasta alcanzar el 50 % del grafo infectado

4.4. Velocidad de difusión siguiendo el modelo epidemiológico SIS

Con el fin de estudiar los efectos de un modelo ligeramente más complejo y realista como es el SIS se ha incluido la variable fija τ mediante la cual se calcula la probabilidad de que el nodo olvide la información como $\delta = \frac{\beta}{\tau}$.

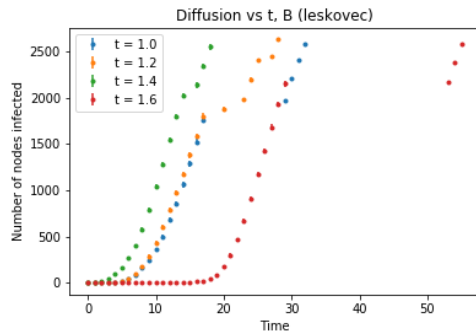


(a) Velocidad de difusión variando τ

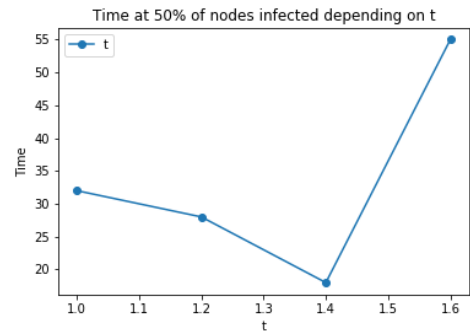


(b) Tiempo empleado para alcanzar el 50 % de los nodos infectados

Figura 4.9: Velocidad de difusión en Barabási según el modelo SIS hasta alcanzar el 50 % del grafo



(a) Velocidad de difusión variando τ



(b) Tiempo empleado para alcanzar el 50 % de los nodos infectados

Figura 4.10: Velocidad de difusión en Leskovec según el modelo SIS hasta alcanzar el 50 % del grafo

En este caso, al aumentar la probabilidad de olvidar la información mediante la variación de la variable τ de 1 a 1.6 se puede observar un incremento en el tiempo de difusión. Esto se debe a que la pérdida de la información además implica que el nodo no intervenga en su propagación a otros, lo que ralentiza el proceso de difusión. Sin embargo, al igual que en el modelo SI, la relación no es completamente lineal debido a la estructura dispersa de los grafos. En concreto se observa que la relación se hace más obvia en las figuras 4.13(b), 4.14(b), 4.15(b) y 4.16(b) al eliminar aquellos nodos con mayor grado y coeficiente de *clustering* ya que esto implica eliminar en gran medida aquellos

nodos que presentan mayores diferencias estructurales.

4.5. Control de la difusión

4.5.1. Efecto de *fact-checking*

Con el fin de simular el proceso de difusión incluyendo la posibilidad de que los nodos verifiquen la información, el modelo de desinformación [16] incluye la variable p_{verify} , que expresa la probabilidad de que el nodo compruebe que la información es falsa y no lo comparta. En este modelo se plantea también la existencia de una constante α que indica la credibilidad de la noticia.

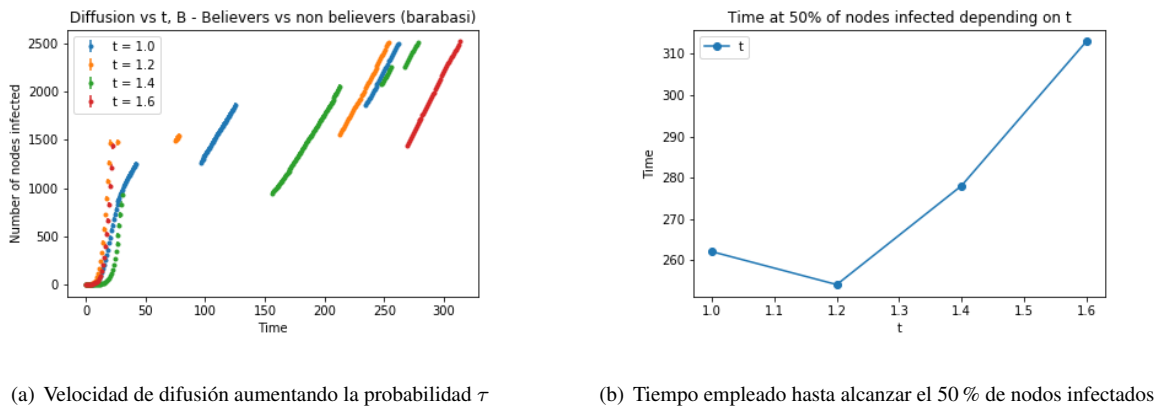


Figura 4.11: Velocidad de difusión en Barabási empleando el modelo de difusión de desinformación

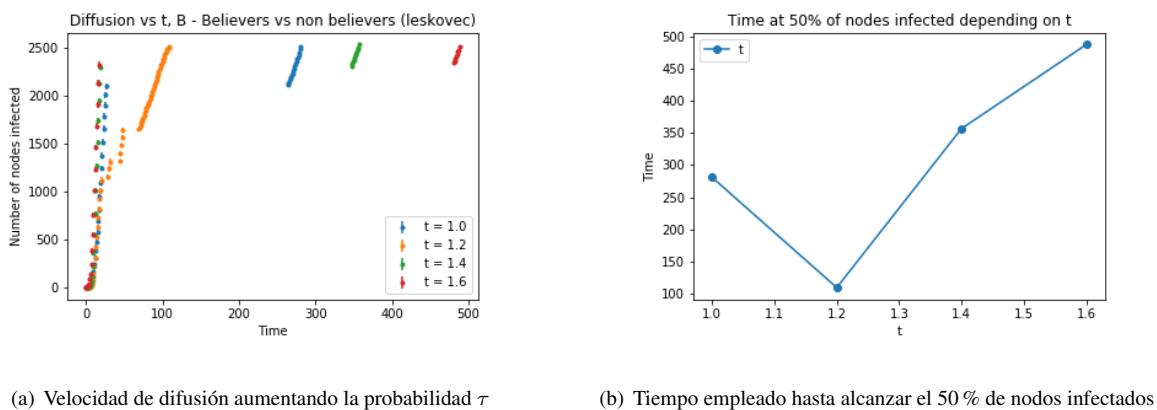


Figura 4.12: Velocidad de difusión en Leskovec empleando el modelo de difusión de desinformación

La probabilidad p_{verify} afecta sustancialmente a la velocidad de difusión ya que el conocimiento de que la noticia es falsa también se propaga, convirtiendo a los nodos vecinos en *fact-checkers* que por tanto no solo no difunden la información sino que contribuyen a su verificación. De hecho, existe un

umbral a partir del cual la noticia falsa se eliminará del grafo debido al proceso de verificación, este es:

$$p_{verify} \geq \frac{2\alpha}{1-\alpha} \cdot p_{forget} \quad (4.7)$$

Sin embargo, en nuestro modelo se ha tomado un parámetro de credibilidad $\alpha = 1$ tal que la noticia siempre se propaga si se cumple $P_b = 1 - (1 - \beta)^{n_b}$. Esto implica que $p_{verify} \rightarrow \infty$, por lo que en nuestro caso el rumor nunca se eliminará del grafo, ralentizándose su velocidad de difusión infinitamente. En este caso, dado que la probabilidad de difusión β se ha fijado a 0,02, a la probabilidad de comprobación del rumor p_{verify} se le ha dado valor 0,01, de forma que sea doblemente probable que un usuario difunda el rumor a que lo verifique.

Se puede comprobar que la verificación de la información resulta una forma de controlar su difusión. Como se muestra en las figuras 4.11(a) y 4.12(a), el contraste de un rumor y su consiguiente retención ralentiza su difusión significativamente.

4.5.2. Eliminación de los nodos con más vecinos

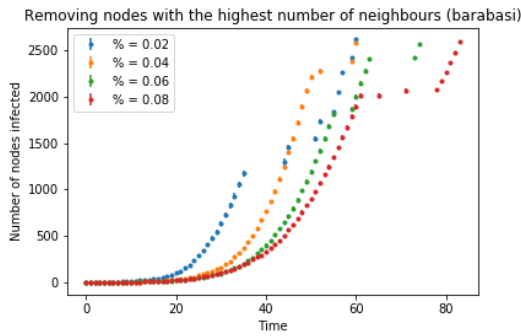
A fin de estudiar el impacto de los nodos con más vecinos sobre la velocidad de difusión se han representado las figuras 4.13(a) y 4.14(a) en las que se ha eliminado progresivamente del 2 % al 8 % de estos nodos. Al eliminarlos la velocidad disminuye sustancialmente lo que prueba que estos juegan un papel fundamental en la difusión. Estos nodos cuentan con una probabilidad de infección potencial mayor al tener más vecinos, ya que implica que el conjunto de vecinos infectados es presumiblemente mayor que en un nodo con un grado menor. Además, la eliminación de estos nodos puede implicar que algunos queden desconectados, incomunicándolos del resto del grafo e impidiendo su contagio.

No obstante, se aprecia una diferencia en la ralentización de la difusión en ambos modelos, teniendo una mayor repercusión en el grafo generado con el modelo de Leskovec. Esto tiene su explicación en la estructura de ambos grafos, ya que como se muestra en las figuras 4.1(a) y 4.1(b), los nodos con más vecinos en este modelo alcanzan grado 800, mientras que en Barabási el mayor grado registrado es 200. Además, en este modelo los nodos con grados superiores a 10 representan solo un 30 % del total, por lo que una vez se elimina hasta el 8 % de este conjunto, estos solo encarnan un 22 % del grafo, mientras que en Barabási prevalecería un 35 % de nodos con grado superior a 10.

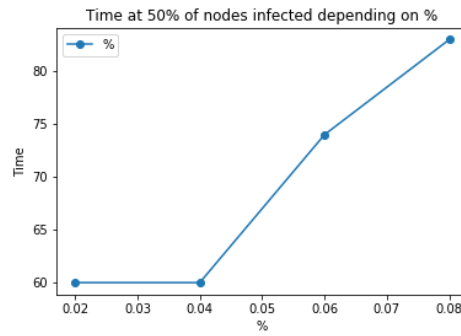
Asimismo, aunque exista una diferencia sustancial entre los grados superiores de cada modelo, al eliminar un 8 % de estos nodos en Leskovec, el grado más elevado existente disminuye considerablemente hasta ser muy cercano al de Barabási. Esto se debe a que el modelo de Leskovec genera grafos con una *power law* más notable, lo que significa que sus nodos tienen grados más dispares entre sí, existiendo un pequeño conjunto de nodos con grados muy elevados así como una gran concentración de nodos con grado 2. Es más, al eliminar este pequeño conjunto de nodos, cobran aún más impor-

tancia los grados con más ocurrencias, que en el caso de Leskovec son los grados 2 y 3, mientras que en Barabási es el grado 7, por lo que se puede verificar que esta eliminación perjudica más al primer modelo.

Igualmente, cabe destacar que al eliminar estos nodos, las gráficas 4.13(b) y 4.14(b) muestran una relación lineal más consistente entre la probabilidad de olvidar la información δ y el tiempo que se necesita para llegar al 50 % de la infección del grafo. Puesto que se han eliminado los nodos con más vecinos, los nodos cuyos grados son más frecuentes suponen una parte más significativa del grafo, ya que estos nunca formarán parte del conjunto eliminado al ser nodos con grados inferiores (10 o menos). Esto supone que la estructura del grafo sea más homogénea, por lo que el número de vecinos infectados no desempeña un rol tan decisivo como lo hacía antes de la eliminación de estos nodos. Siendo la distribución de grados más uniforme, el aumento de la variable τ es un factor más distintivo en lo referente a la velocidad de propagación, ya que tiene un impacto mayor en el cálculo de la probabilidad con la que un nodo cambia de estado.

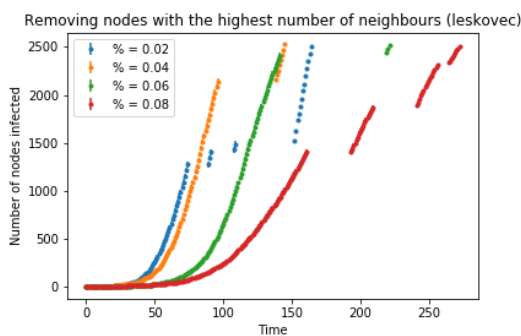


(a) Velocidad de difusión aumentando el número de nodos eliminados en el modelo de Barabási

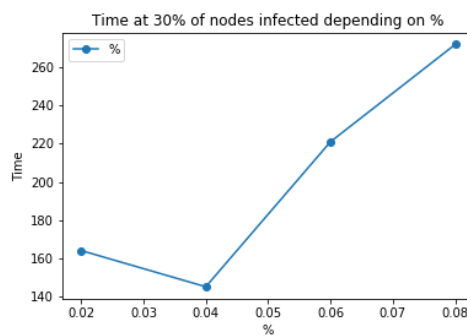


(b) Tiempo para alcanzar el 50 % de nodos infectados

Figura 4.13: Velocidad de difusión eliminando los nodos con más vecinos en el modelo de Barabási



(a) Velocidad de difusión aumentando el número de nodos eliminados en el modelo de Leskovec

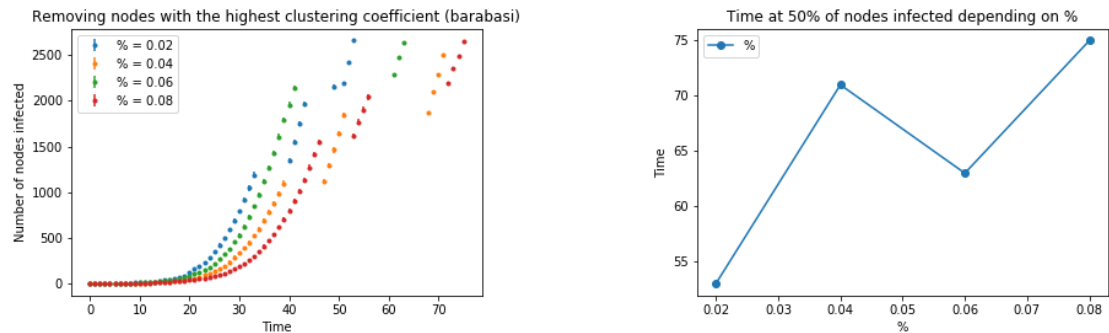


(b) Tiempo para alcanzar el 50 % de nodos infectados

Figura 4.14: Velocidad de difusión eliminando los nodos con más vecinos en el modelo de Leskovec

4.5.3. Eliminación de los nodos con más *clustering*

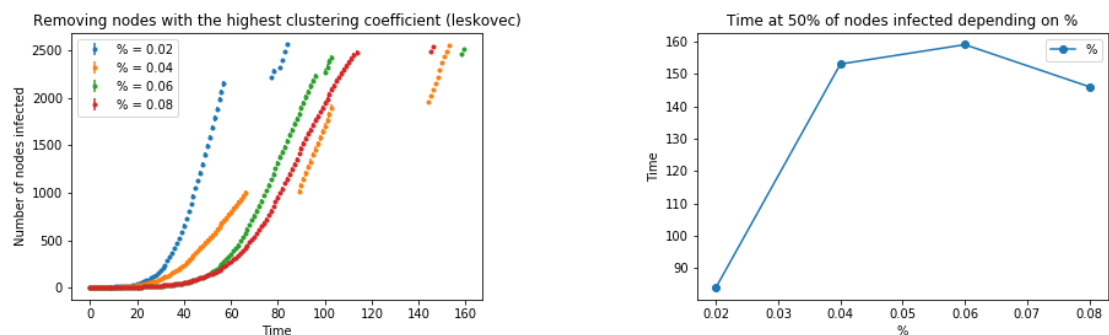
Al igual que eliminar los nodos con más vecinos en cada grafo, los nodos con mayor *clustering* también tienen efectos visibles en la disminución de la velocidad de difusión, provocando tiempos más elevados.



(a) Velocidad de difusión aumentando el número de nodos eliminados en el modelo de Barabási

(b) Tiempo para alcanzar el 50 % de nodos infectados

Figura 4.15: Velocidad de difusión eliminando los nodos con más *clustering* en Barabási



(a) Velocidad de difusión aumentando el número de nodos eliminados en el modelo de Leskovec

(b) Tiempo para alcanzar el 50 % de nodos infectados

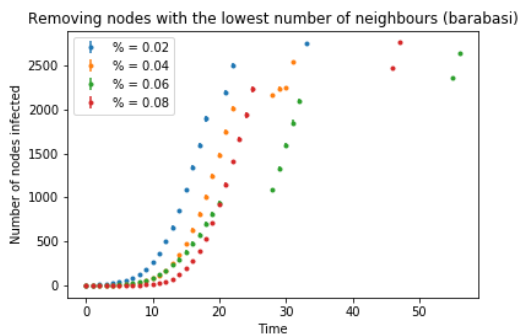
Figura 4.16: Velocidad de difusión eliminando los nodos con más *clustering* en Leskovec

Esta modificación tiene un efecto mayor en el modelo de Leskovec de nuevo, como se puede ver en las gráficas 4.15(a) y 4.16(a), ya que el *clustering* es una propiedad más notable en los grafos generados por este modelo. En la figura 4.3(a) se expone que en el modelo de Barabási el 91 % de los nodos tienen un coeficiente de *clustering* inferior a 0,1 y los mayores coeficientes registrados son de 0,6, mientras que en el modelo de Leskovec la diferencia entre los coeficientes más frecuentes (entre 0 y 1) y los más altos (4) es mucho mayor. Debido a esto, eliminar hasta el 9 % restante de los nodos con mayor frecuencia en Barabási no tiene un impacto tan elevado como eliminar este porcentaje en el grafo de Leskovec, cuyo coeficiente es mucho mayor de forma generalizada.

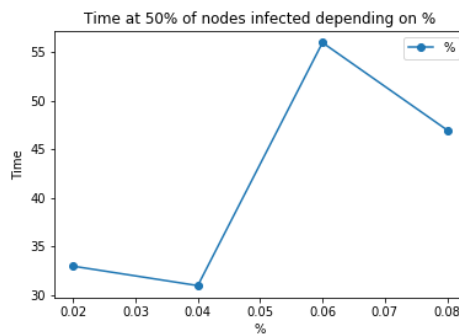
4.5.4. Eliminación de los nodos con menos vecinos y menos *clustering*

Sin embargo, no basta con eliminar cualquier conjunto de nodos, con objetivo de demostrar que no es el hecho de eliminar nodos lo que varía la velocidad de difusión, se han observado los efectos de eliminar los nodos con menos vecinos y menor *clustering* del grafo respectivamente.

A pesar del impacto que tienen los nodos con más vecinos y *clustering* sobre la velocidad de difusión, en la figura 4.17(a) se hace visible que la supresión de los nodos con menos vecinos no afecta apenas a esta velocidad. El efecto observado al eliminar aquellos con menor *clustering* ha sido el mismo por lo que se ha considerado redundante incluir estas figuras.



(a) Velocidad de difusión aumentando el número de nodos eliminados en el modelo de Barabási



(b) Tiempo para alcanzar el 50 % de nodos infectados

Figura 4.17: Velocidad de difusión eliminando los nodos con menos vecinos en Barabási

Esto se debe al fenómeno de las redes de mundo pequeño [21]. Estas redes se caracterizan esencialmente porque aunque la mayoría de nodos no son vecinos entre sí, en general se puede alcanzar cualquier nodo desde cualquier punto de la red empleando una cantidad muy reducida de saltos para ello. Otras características propias de estas redes son la presencia de *cliques*; o subgrafos cuyos nodos están altamente conectados entre sí, lo que se traduce en un alto coeficiente de *clustering*; y la existencia de *hubs*, nodos con un elevado número de conexiones y por lo tanto un alto grado. Debido a estas propiedades, los grafos presentados no solo se tratan de redes de mundo pequeño, sino que, R. Cohen y Havlin mostraron en sus investigaciones sobre las propiedades estructurales de las redes libres de escala que estas tienen cualidades propias de redes de mundo ultra-pequeño.

Es debido a la estructura de estas redes que la eliminación de un nodo aleatorio, y en especial de un nodo que no forma parte del conjunto de *hubs* del grafo, no supone una gran variación en el camino más corto entre dos nodos. De esta forma, dado que la mayoría de caminos más cortos entre nodos atraviesan *hubs*, la exclusión de nodos con grado o *clustering* bajos no tendrá una repercusión notable en la difusión a través de la red. De hecho, incrementar el porcentaje a eliminar de este conjunto de nodos tampoco tendría un impacto remarcable mientras este no alcanzase al conjunto minoritario de nodos que actúan como *hubs* [22].

CONCLUSIONES

Como se ha podido observar en los últimos meses, la información se propaga incluso más rápido que los virus biológicos [23], empleando como medio de transmisión por excelencia las redes sociales. Esta transmisión está altamente relacionada con el tipo de contenido que se difunde, ya que como se ha mencionado anteriormente [2], la desinformación se propaga a una velocidad mucho mayor que la información verdadera. En este estudio se ha podido comprobar que la rapidez con la que se difunde esta información depende en gran medida de ciertos elementos propios de las estructuras de las redes sociales, a saber: el número de interconexiones que existen entre los vecinos de un usuario o clustering y especialmente el número de contactos o seguidores de los nodos, es decir, sus vecinos en el grafo.

Esta observación es especialmente importante en el estudio del control de la propagación de desinformación, ya que al comparar el efecto del uso de metodologías de *fact-checking* con la eliminación de ciertas características estructurales de las redes, se pueden advertir efectos similares en cuanto a la ralentización de la transmisión.

Es evidente que es imposible eliminar la difusión de noticias falsas en su totalidad debido al atractivo de su contenido para los usuarios, tanto por su carácter novedoso y su contenido generalmente llamativo como por el deseo innato de ser conocedor y fuente de las últimas noticias. No obstante, dado que las noticias falsas de índole político se encuentran entre los grupos de información con mayor viralidad, es natural que exista cierta preocupación con su propagación de forma masiva debido a su influencia. Por ello, es de gran interés actualmente identificar ciertos puntos de la red que juegan un papel fundamental en la transmisión. Aunque como se ha mencionado este se trata de un problema NP completo, las aproximaciones empleadas han dado resultados esperanzadores. En este estudio se han obtenido resultados muy similares a aquellos del modelo de verificación de desinformación al eliminar los nodos con más vecinos. Esto implica que una eliminación de un porcentaje tan pequeño como el 2 % de aquellos nodos con mayor grado tiene un impacto muy significativo sobre la rapidez de la difusión y como consecuencia, su alcance. Evidentemente la eliminación de estos usuarios de las redes sociales no es viable en el contexto de la vida real, sin embargo estos resultados se pueden traducir en métodos más factibles tales como el control y supervisión de los nodos que presentan este nivel de influencia en la red.

Una medida que ejemplifica este concepto se ha visto ejecutada por *Twitter* por primera vez recientemente en un *tweet* en el que el presidente de los Estados Unidos, Donald Trump, hablaba del voto por correo como un método fraudulento de votación [24]. Al poco tiempo de su publicación, este *tweet* fue acompañado de una advertencia agregada por *Twitter* de que su contenido podía contener información inexacta, junto con un link a otras fuentes que aportaban más información sobre este tipo de votación. Esta es una muestra de que la implantación de este tipo de medidas es posible, pero también plantea un interrogante moral, ya que al fin y al cabo quienes verifican la información son personas y como es natural, pueden ser parciales o tener ciertos prejuicios a la hora de validar la información.

Ciertamente en algunos casos la veracidad del contenido será fácilmente desacreditada, sin embargo siempre cabe la posibilidad de que exista información cuya ratificación sea más compleja, y entre en un área más subjetiva. No solo eso, sino que uno también puede discutir que métodos como este atentan contra la libertad de expresión de los individuos. Este es un tema muy complejo que probablemente será muy debatido en los próximos años debido a la rápida expansión de las redes sociales y la necesidad de definir ciertos controles sobre estas.

BIBLIOGRAFÍA

- [1] S. Kemp, "Digital 2020: Global digital overview," 2020. <https://datareportal.com/reports/digital-2020-global-digital-overview> [Último acceso: 15-05-2020].
- [2] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018. (Descargar).
- [3] J. W. C. Marc Fisher and P. Hermann, "Pizzagate: From rumor, to hashtag, to gunfire in d.c.," 2016. https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html [Último acceso: 10-05-2020].
- [4] P. Beaumont, "The truth about twitter, facebook and the uprisings in the arab world," 2011. <https://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya> [Último acceso: 08-05-2020].
- [5] A. Majmundar, J.-P. Allem, T. Boley Cruz, and J. B. Unger, "The why we retweet scale," *PLOS ONE*, vol. 13, no. 10, pp. 1–12, 2018. (Descargar).
- [6] P. Kafka, "The pandemic is driving media consumption way up. but ad sales are falling apart.," 2020. <https://www.vox.com/recode/2020/3/24/21192311/coronavirus-pandemic-media-advertising-twitter-new-york-times> [Último acceso: 11-05-2020].
- [7] D. A. Ceren Budak and A. E. Abbadi, "Limiting the spread of misinformation in social networks," *WWW 2011 – Session: Information Credibility*, vol. 1, no. 1, pp. 665–674, 2011. (Descargar).
- [8] U. N. del Sur (Argentina), "Análisis de redes sociales," 2013. <http://ars-uns.blogspot.com/2013/10/coeficiente-de-agrupamiento.html> [Último acceso: 30-06-2020].
- [9] A. Mahanti, N. Carlsson, M. Arlitt, and C. Williamson, "A tale of the tails: Power-laws in internet measurements," *Network, IEEE*, vol. 27, pp. 59–64, 2013. (Descargar).
- [10] N. Mishra, R. Schreiber, I. Stanton, and R. Tarjan, "Clustering social networks," pp. 56–67, 2007. (Descargar).
- [11] D. Choi, S. Chun, H. Oh, J. Han, and T. T. Kwon, "Rumor propagation is amplified by echo chambers in social media.," *Scientific Reports*, vol. 10, no. 310, pp. 1–10, 2020. (Descargar).
- [12] P. Erdős and A. Rényi, "On random graphs. i," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959. (Descargar).
- [13] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999. (Descargar).
- [14] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 462–470, 2008. (Descargar).

- [15] I. V. Management, “Si and sis models,” 2019. <https://idmod.org/docs/malaria/model-si.html> [Último acceso: 15-05-2020].
- [16] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer, “Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks,” 2015. (Descargar).
- [17] G. P. Basharin, A. N. Langville, and V. A. Naumov, “The life and work of a. a. markov.”
- [18] Numerentur.org, “Modelos de markov.” <http://numerentur.org/modelos-de-markov/> [Último acceso: 15-06-2020].
- [19] J. Rocca, “Introduction to markov chains,” 2019. <https://towardsdatascience.com/brief-introduction-to-markov-chains-2c8cab9c98ab> [Último acceso: 15-06-2020].
- [20] I. J. O. Piet Van Mieghem, Member and R. Kooij, “Virus spread in networks,” *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 1–3, 2009. (Descargar).
- [21] D. Watts and S. Strogatz, “Collective dynamics of ‘small-world’ networks.,” *Nature*, vol. 393, pp. 440–442, 1998. (Descargar).
- [22] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function*. Cambridge University Press, 2010.
- [23] L. Hébert-Dufresne, S. V. Scarpino, and J.-G. Young, “Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement,” vol. 16, no. 4, pp. 426–431, 2020. (Descargar).
- [24] K. Conger and D. Alba, “Twitter refutes inaccuracies in trump’s tweets about mail-in voting,” 2020. <https://www.nytimes.com/2020/05/26/technology/twitter-trump-mail-in-ballots.html> [Último acceso: 30-06-2020].

DEFINICIONES

backtracking Técnica algorítmica para resolver problemas de manera recursiva que trata de construir una solución de forma incremental, eliminando aquellas soluciones que no satisfacen las restricciones del problema.

bot Tipo de programa informático autónomo que es capaz de llevar a cabo tareas concretas e imitar el comportamiento humano.

engagement Compromiso que se establece entre una marca y su audiencia en las distintas comunicaciones que producen entre sí.

fake news Tipo de bulo que consiste en un contenido pseudoperiodístico difundido a través de portales de noticias, prensa escrita, radio, televisión y redes sociales y cuyo objetivo es la desinformación.

influencer Persona que destaca en una red social u otro canal de comunicación y expresa opiniones sobre un tema concreto que ejercen una gran influencia sobre muchas personas que la conocen.

phishing Estafa que tiene como objetivo obtener a través de internet datos privados de los usuarios, especialmente para acceder a sus cuentas o datos bancarios.

power law Relación funcional entre dos cantidades, donde un cambio relativo en una cantidad resulta en un cambio relativo proporcional en la otra cantidad, independientemente del tamaño inicial de esas cantidades: una cantidad varía como una potencia de otra.

self-loop Arista que conecta un vértice consigo mismo.

viralidad Modo de transmisión indirecta de la información por la red en un corto espacio de tiempo.

ACRÓNIMOS

NP Nondeterministic Polynomial Time.

SI Susceptible Infected.

SIS Susceptible Infected Susceptible.

